UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS

Faculdade de Ciências Departamento de Biologia

Phylogenomic and population genomic insights on the evolutionary history of Coffee Leaf Rust within the rust fungi

" Documento Definitivo"

Doutoramento em Biologia e Ecologia das Alterações Globais
Especialidade de Biologia do Genoma e Evolução

Diogo Nuno Proença Rico Silva

Tese orientada por:
Doutora Dora Cristina Vicente Batista Lyon de Castro
Prof. Doutor Octávio Fernando de Sousa Salgueiro Godinho Paulo

Documento especialmente elaborado para a obtenção do grau de doutor

2018

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS

Faculdade de Ciências Departamento de Biologia

Phylogenomic and population genomic insights on the evolutionary history of Coffee Leaf Rust within the rust fungi

Doutoramento em Biologia e Ecologia das Alterações Globais
Especialidade de Biologia do Genoma e Evolução

Diogo Nuno Proença Rico Silva

Tese orientada por:
Doutora Dora Cristina Vicente Batista Lyon de Castro
Prof. Doutor Octávio Fernando de Sousa Salgueiro Godinho Paulo

Júri:
Presidente:
- Doutora Maria Manuela Gomes Coelho de Noronha Trancoso, Professora Catedrática e Membro do Conselho Cientifico da Faculdade de Ciências de Lisboa

Vogais:
- Doutor Robert Fraser Park, Professor *Faculty of Science* da *University of Sydney*, (Austrália);
- Doutor Ludwig Krippahl, Professor Auxiliar Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa;
- Doutor Amadeu Mortágua Velho da Maia Soares, Professor Catedrático Departamento de Biologia da Universidade de Aveiro;
- Doutora Maria Helena Mendes da Costa Ferreira Correia de Oliveira, Professora Associada Instituto Superior de Agronomia da Universidade de Lisboa;
- Doutora Maria da Luz da Costa Pereira Mathias, Professora Catedrática Faculdade de Ciências da Universidade de Lisboa;
- Doutora Dora Cristina Vicente Batista Lyon de Castro, Bolseira Pós-doutoramento Faculdade de Ciências da Universidade de Lisboa (Orientadora).

Documento especialmente elaborado para a obtenção do grau de doutor

Fundação para a Ciência e a Tecnologia (FCT) – Bolsa de Doutoramento

(SFRH/BD/86736/2012)

2018

Acknowledgements

Even though the PhD dissertation is mainly regarded as the product of an individual work, I felt that this journey was filled with team-work and collaborative efforts, not only at the scientific and professional levels but also at the personal level. Among these, there are some people that I would like to express my gratitude in particular:

To my advisers, Dr. Dora Batista and Prof. Dr. Octávio S. Paulo for accepting me as a PhD student and for their continuous support and discussions during the entire project.

To Dora, I would like to express my admiration for your tenacity, knowledge of the biological world and hard work when it comes to the scientific enterprise. In particular, I thank you for the "maternal"-like support provided throughout the PhD, which was actually key for maintaining my motivation at several parts of the project, even when you were not aware of it. The PhD project can be incredibly wearing and frustrating at times, and your interventions were instrumental for the successful conclusion of this work.

To Octávio, I would like to thank you for the unwavering words of encouragement, motivation and guidance. Your continuous positive outlook on science was inspirational and our brainstorming sessions were an incredible motivator for doing more and better science. Our discussions were exceptionally stimulating scientifically and I only wish they could have been more frequent.

To the team at the Centro de Investigação das Ferrugens do Cafeeiro (CIFC), in particular Eng. Vítor Várzea, Dr. Maria do Céu and Dr. Paula Pereira, I thank you for receiving and supporting me and for the words of encouragement. To the CoBiG^2 group, I would also like to thank for all the discussions and support.

To Dr. João Carriço and Prof. Dr. Mário Ramirez, at the Instituto de Medicina Molecular (IMM), I am grateful for giving me the opportunity to work on a new and exciting project while also providing all the necessary conditions to finish the PhD.

To Tiago F. Jesus, I would like to report! my gratitude for those bandages, for carrying me through those hellish rifts, for having my back against unnatural hordes, for the numerous revives and for the house and life we built together. In Minecraft. I am also grateful for the long and continuous friendship, for the emotional support

and for all the scientific discussions. You will always be my number 1 beta tester, though I'll never forgive your lack of intervention when I was at the altar.

To Bruno Novais, I am grateful for breaking a leg for me, for all the healing and spotting, the broken promises and, most importantly, the stolen chickens. Moreover, your tongue work was both unforgettable and awestrucking. Thank you for the friendship and all the hilarious moments. I hope we can resume the hand in hand bike riding soon.

To Fernano Alves, dear sir, Bro, illustrious gentleman, you were a pillar of support and friendship during and beyond this period. You have showed me the way many times, all for the glory of our equine overlords. Also, thank you for the collaboration in removing the pain of Perl and SQL from my to do list. This was a sacrifice than I am not sure I'll ever be able to repay, but for which I am forever grateful.

To João Caetano, our most beloved and mighty DM, I am grateful for expanding my music tastes (Am I really..?). You have bestowed upon me many wondrous things, like your knowledge on SQL databases and other programming affairs, your friendship and hilarious life-stories, Artur Zé, and Yolandi Vesser.

To Francisco Pina-Martins, my dear open-* advocate, Linux wizard and friend. All those sessions of coding, tweaking with stuff and discussing science were a blast and something that I miss. I really hope that ours paths cross again so that we can continue with those sessions. In any case, we'll always have Stallman's "talk".

To the geek squad, with illustrious members like Diogo Simões and Daniel Alves, I thank you for your friendship, pure unfiltered geekness, tabletop galore, and so much hilarious and fond memories. Our sessions were a bliss during the hard times of the PhD and a much needed sanity boost.

However, these words would not be being written, if not for my wife, parents and grandmother.

To my parents, thank you for the patience and comprehension, for always believing in me and for providing all the necessary conditions to reach where I am. In so many ways, your unwavering support was fundamental to the successful conclusion of this project. You have set the parenthood role model which one day I aspired to be.

To Yana, my wife and best friend with whom I shared everything, I extend my deepest gratitude. You were there for every moment of joy, accomplishment and happiness, but most importantly, you supported me in every moment of frustration, self-doubt and sadness. You were a pillar of sanity and my most powerful motivator. My main source of happiness, even when everything around was exploding in poop. Thank you for all the unique and irreplaceable moments, without which I would not reach this far, but most of all, thank you for your love.

Contents

Abstract **xi**

Resumo **xiii**

List of Figures **xvii**

List of Tables **xxiii**

1 Introduction **1**

1.1 The promise of the genomics era 1

1.1.1 Phylogenomics 3

1.1.2 Population genomics and RAD sequencing 7

1.1.3 The rise of bioinformatics 11

1.1.4 The genomic revolution for fungal plant pathogens 13

1.2 The rust fungi 15

1.2.1 Life-style, life-cycle and infection 17

1.2.2 Pathogenicity and host interaction 18

1.2.3 Genomic features 20

1.2.4 *Hemileia vastatrix* - Coffee Leaf Rust 21

1.3 Objectives 25

1.4 References 26

2 Genomic patterns of positive selection at the origin of rust fungi **39**

2.1 Abstract 39

2.2 Introduction 40

2.3 Materials and Methods 43

2.3.1 Genomic and EST data 43

2.3.2 Processing of EST data 44

2.3.3 Ortholog search strategy 44

2.3.4 Sequence alignment and filtering 45
2.3.5 Data set assembly 46
2.3.6 Phylogenomic reconstruction 47
2.3.7 Detection of positive selection at the origin of the Pucciniales 47
2.3.8 Assessment of gene molecular rates 48
2.3.9 Functional annotation 49
2.4 Results 50
2.4.1 Assembly of phylogenomic data sets 50
2.4.2 Evolutionary history of the Pucciniales 54
2.4.3 Episodic positive selection at the origin of the rust fungi . . . 54
2.4.4 Estimation of relative evolutionary rates 56
2.4.5 Functional annotation and enrichment 57
2.5 Discussion 60
2.5.1 Genome-wide scan for positive selection 60
2.5.2 Assessing the evolutionary rate on the origin of Pucciniales . 63
2.5.3 Functional enrichment analyses 65
2.6 Supplementary information 70
2.6.1 Positive selection on conserved amino acid sites 70
2.7 References 76

3 Population genomic footprints of host adaptation, introgression and recombination in Coffee Leaf Rust 83

3.1 Abstract 83
3.2 Introduction 84
3.3 Materials and methods 87
3.3.1 Fungal material, sample preparation and RAD sequencing . 87
3.3.2 RADseq assembly strategy and SNP calling 88
3.3.3 Phylogenetic analyses 88
3.3.4 Evaluation of *H. vastatrix* genetic structure 89
3.3.5 Introgression assessment 89
3.3.6 Recombination and linkage disequilibrium 90
3.3.7 Population dynamics of *H. vastatrix* infecting tetraploid hosts 91
3.4 Results 91
3.4.1 RAD-Seq data assembly and quality control 91

3.4.2 Phylogenetic analysis 92
3.4.3 Population structuring of *H. vastatrix* 94
3.4.4 Genetic diversity 95
3.4.5 Investigating introgression between *H. vastatrix* groups . . . 96
3.4.6 Linkage disequilibrium and recombination 98
3.4.7 Population dynamics of *H. vastatrix* in tetraploid hosts 99
3.5 Discussion 100
3.5.1 Revealing *H. vastatrix* as a potential cryptic species complex with host specialization 100
3.5.2 Investigating the presence of introgression between groups of *H. vastatrix* infecting diploid and tetraploid hosts 102
3.5.3 The emergence and evolutionary history of the C3 group . . 104
3.5.4 Conclusion and remarks on CLR disease management . . . 107
3.6 Supplementary data 108
3.6.1 Assembly strategy 108
3.7 References 113

4 TriFusion: Streamlining phylogenomic data gathering, processing and visualization 119
4.1 Abstract 119
4.2 Introduction 120
4.3 Description 121
4.3.1 Orthology: Search and explore orthologs 122
4.3.2 Process: Manipulation and processing of alignments 123
4.3.3 Statistics: Statistics and visualization of alignment data . . . 127
4.3.4 Command line versions 128
4.4 Algorithm description and performance 128
4.4.1 Alignment parsers 129
4.4.2 Data retrieval 135
4.4.3 Main operations 140
4.4.4 Secondary operations 146
4.4.5 Writers 150
4.5 Continuous integration, unit testing and code quality 155
4.5.1 Continuous integration 156

4.5.2 Unit/integration tests and code quality 157
4.6 Benchmarks and Biological Examples 164
4.7 Availability 166
4.8 Supplementary data 167
4.8.1 Benchmarks with third-party software 167
4.9 References 173

5 Final Remarks **175**
5.1 Conclusions and main contributions 176
5.1.1 Phylogenomics 176
5.1.2 Population genomics 179
5.1.3 Software development 180
5.2 Future perspectives 181
5.3 References 183

A Appendix **185**
A.1 Genomic patterns of positive selection at the origin of rust fungi . . . 185
A.1.1 Tables 185
A.2 Population genomic footprints of host adaptation, introgression and recombination in Coffee Leaf Rust 210
A.2.1 Tables 210
A.2.2 Figures 218
A.3 TriFusion: Streamlining phylogenomic data gathering, processing and visualization 221
A.3.1 Tables 221

Abstract

Fungi are currently responsible for more than 30% of the emerging diseases worldwide and rust fungi (Pucciniales, Basidiomycota) are one of the most destructive groups of plant pathogens. In this thesis, two genomic approaches were pursued to further our knowledge on these pathogenic fungi at the macro-evolutionary level, using phylogenomics, and micro-evolutionary level, using population genomics. At the macro-evolutionary level, a phylogenomics pipeline was developed with the aim of investigating the role of positive selection on the origin of the rusts, particularly related to their obligate biotrophic life-style and pathogenicity. With up to 30% of the *ca.* 1000 screened genes showing a signal of positive selection, these results revealed a pervasive role of natural selection on the origin of this fungal group, with an enrichment of functional classes involved in nutrient uptake and secondary metabolites. Furthermore, positive selection was detected on conserved amino acid sites revealing an unexpected but potentially important role of natural selection on codon usage preferences. At the micro-evolutionary level, the focus was shifted to the coffee rust, *Hemileia vastatrix*, which is the causal agent of leaf rust disease and the main threat to Arabic coffee production worldwide. Using RAD sequencing to produce thousands of informative SNPs for a broad and unique sampling of this species, the aim was to investigate its evolutionary history and translate population genomic insights into recommendations for disease control. The results of this work overturned most of the preconceptions about the pathogen by revealing that instead of a single unstructured and large population, *H. vastatrix* is most likely a complex of cryptic species with marked host specialization. Moreover, genomic signatures of hybridization and introgression occurring between these lineages were uncovered, raising the possibility that virulence factors may be quickly exchanged. The most recent “domesticated” lineage infects exclusively the most important coffee species and SNP linkage analyses revealed the presence of recombination among isolates

that were previously thought to be clonal. Altogether, these results considerably raise the evolutionary potential of this pathogen to overcome disease control measures in coffee crops. To undertake most of the tasks in this project, a new computational application called TriFusion was developed to streamline the gathering, processing and visualization of big genomic data.

Resumo

Os fungos são responsáveis por mais de 30% das doenças emergentes de plantas em todo o mundo, incluindo algumas das doenças mais destrutivas que se conhecem. De um modo alarmante, este número terá tendência a aumentar e apesar de todos os avanços tecnológicos, as acções antropogénicas têm contribuído fortemente para esta ameaça com a globalização das trocas comerciais e as práticas agrícolas modernas. De forma a mitigar e reverter esta tendência, a inclusão de estudos eco-evolutivos no delineamento de medidas de controlo reveste-se de uma importância acrescida, devido ao seu potencial para aprofundar o nosso conhecimento sobre os mecanismos da patogenicidade e história evolutiva de grupos de espécies fúngicas. Na presente tese, foram utilizadas abordagens genómicas inovadoras com o intuito de aprofundar o nosso conhecimento sobre um importante grupo de patógenos fúngicos, as ferrugens (Pucciniales, Basisidiomycota). Com o aparecimento e democratização das tecnologias de sequenciação de alto débito, tem sido possível trabalhar com DNA nuclear em larga escala, mesmo em espécies não modelo e sem recursos genómicos prévios. O aumento da acessibilidade em termos financeiros a estas tecnologias tem permitido o desenvolvimento de estudos, que visam questões ecológicas e evolutivas, os quais eram impensáveis há apenas alguns anos atrás. O trabalho desenvolvido nesta tese representa um exemplo disso mesmo. Do ponto de vista macro-evolutivo, foi avaliado o papel da selecção positiva ao nível genómico no ramo filogenético que deu origem às ferrugens, usando ferramentas e abordagens filogenómicas. Do ponto de vista micro-evolutivo, foi efectuado um estudo de genómica populacional de uma espécie importante de ferrugens que causa, a nível mundial, a ferrugem alaranjada do cafeeiro, a *Hemileia vastatrix*.

Na vertente macro-evolutiva, foi tirado partido da disponibilidade de dezenas de genomas e dados de "Expressed Sequence Tags" (EST) para várias espécies de Basidiomicetos. A partir deste conjunto de dados, foi desenvolvida uma "pipeline" filogenómica capaz de encontrar o maior número de genes ortólogos de cópia única entre um conjunto de espécies de Basidiomicetos, incluindo vários genomas e dados de ESTs de ferrugens. Encontrado este conjunto de genes de elevada qualidade, foi construída outra "pipeline" para medir e testar o papel da selecção positiva no ramo filogenético que dá origem às ferrugens. Pretendia-se deste modo avaliar o papel da selecção natural na origem deste grupo de patógenos focando em genes conservados e partilhados com outras espécies do filo. Apesar da conservação, 20-30% destes genes demonstraram a presença de assinaturas de selecção positiva, revelando a importância do papel da selecção natural no processo evolutivo. Análises funcionais revelaram ainda que as classes de genes envolvidos na obtenção e transporte de nutrientes, assim como na biossíntese de metabolitos secundários, estavam substancialmente enriquecidas em genes sob selecção. Estes resultados explicam as profundas alterações que as ferrugens terão sofrido ao nível da aquisição de nutrientes quando transitaram para um estilo de vida obrigatoriamente biotrófico e fornecem um ponto de partida para estudos funcionais futuros. No entanto, o resultado mais intrigante deste estudo foi a descoberta de um sinal de selecção positiva considerável (15-24%) a actuar sobre aminoácidos conservados em todas ou quase todas as espécies estudadas. Este aparente paradoxo ocorreu em aminoácidos como a Serina que são codificados por um conjunto alargado de codões, como "TCN" ou "AGY", em que a transição entre eles implica 2 a 3 mutações não sinónimas. Análises subsequentes demonstraram que a utilização de codões diferia substancialmente entre ferrugens e outras espécies de Basidiomicetos. Várias hipóteses foram formuladas para explicar este fenómeno, mas a actuação directa da selecção natural na escolha de codões específicos devido à existência de alguma vantagem adaptativa apresentou-se como a mais provável e interessante.

Na vertente micro-evolutiva, foi investigada a história evolutiva de uma importante espécie de ferrugem que infecta especificamente o cafeeiro, incluindo várias espécies em todo o mundo, a *H. vastatrix*. Para tal, foi utilizada uma técnica de

sequenciação denominada de “RADseq”, para obter um painel de milhares de polimorfismos nucleotídicos únicos (“SNP”) ao longo de todo o genoma e com uma amostragem única e geograficamente alargada de *H. vastatrix*. Na etapa da montagem e genotipagem dos SNPs, uma das mais complicadas do ponto de vista técnico, foi utilizada uma abordagem de replicados técnicos de forma a minimizar o erro na genotipagem e maximizar o número de marcadores moleculares informativos. Este conjunto de dados permitiu desvendar que *H. vastratrix* não era uma espécie única e não estruturada, mas sim um potencial complexo de espécies crípticas com uma acentuada especialização a nível do hospedeiro, agrupando três linhagens divergentes. Duas das linhagens mais basais e antigas foram amostradas em espécies de cafeeiros diploides, como *Coffea canephora*. Estas demonstraram ser capazes de infectar quase exclusivamente estes hospedeiros, com poucas excepções onde produziram sintomas ligeiros em hospedeiro tetraploides. Por outro lado, a linhagem de maior dimensão infectava exclusivamente cafeeiros tetraploides, como *C. arabica* e híbridos inter-específicos, os quais representam os cafeeiros mais importantes do ponto de vista económico. A análise filogenética demonstrou que esta poderá ser uma linhagem domesticada muito mais recente do que as restantes e altamente adaptada a ecossistemas agrícolas, que terá resultado de um salto de hospedeiro, provavelmente a partir de cafeeiros diploides. Além da especialização das linhagens no que se refere ao hospedeiro, os dados genómicos revelaram ainda a presença de hibridação e introgressão entre linhagens. Esta observação levanta a possibilidade preocupante de que estes grupos possam trocar factores de virulência entre eles, e com isso suplantar rapidamente as variedades de cafeeiros resistentes que são introduzidas no campo. Finalmente, foi ainda detectada a presença de recombinação nesta nova linhagem “domesticada”, revelando assim o seu elevado potencial evolutivo para ultrapassar as medidas de controlo impostas. Estes resultados representaram uma mudança de paradigma na nossa compreensão do agente patogénico e da sua epidemiologia, e terão um grande impacto na futura investigação do fungo e no desenvolvimento de medidas de controlo para a doença.

Por fim, o trabalho desenvolvido durante esta tese teve uma forte componente bioinformática que foi essencial para a execução das várias tarefas. Como resultado

disso, foi desenvolvida uma aplicação computational completa chamada TriFusion que teve como objectivo simplificar a obtenção, processamento e visualização de grandes matrizes de dados. Esta aplicação foi desenvolvida com uma interface gráfica e por linha de comandos, oferecendo uma elevada eficiência e rapidez no processamento de conjuntos de dados genómicos. Esta permite a qualquer utilizador realizar tarefas complexas e laboriosas em estudos de filogenómica e genómica populacional, independentemente da sua experiência em bioinformática. Por exemplo, em apenas alguns segundos, o utilizador é capaz de executar uma "pipeline" para obter genes ortólogos a partir de múltiplos genomas completos. O TriFusion é também capaz de concatenar, converter, filtrar e colapsar milhares de alinhamentos com centenas de espécies em simultâneo e preparar as matrizes para vários programas de análise a jusante. Integrado na própria aplicação está também um sistema de construção e visualização de gráficos interactivos para os utilizadores poderem explorar e exportar facilmente os seus resultados. Para além do TriFusion possuir um conjunto de funcionalidades únicas, supera todas as alternativas em termos de desempenho para as tarefas mais comuns e ainda apresenta uma interface intuitiva e de fácil utilização.

List of Figures

1.1 Overview of RAD sequencing library preparation and sequencing. Retrieved from https://www.floragenex.com/rad-seq/. 10

1.2 Dynamics of bioinformatics related publications over the past four decades with the quoted keyword retrieved from scientific publications in PubMed. Adapted from Atwood et al. (2015). 12

1.3 Disease alerts in the ProMED database for pathogenic fungi of animals and plants. Retrieved from Fisher et al. (2012). 14

1.4 Examples of symptomatic (sporulating) infections of rust fungi. **a)** *Puccinia graminis* on wheat, **b)** *Melampsora larici-populina* on poplar, **c)** *Uromyces viciae-fabae* on faba bean and **d)** *Hemileia vastatrix* on coffee. 16

2.1 Schematic representation of the different foreground branches being considered in the basidioPAML and basidioPAML_Hv data sets. 48

2.2 Basidiomycota phylogenetic tree. Maximum likelihood tree illustrating the evolutionary relationships among 67 Basidiomycota and Ascomycota species. Names on the right of the figure correspond to the taxonomic order of the respective highlighted taxa. Within the Pucciniales, the three sub-orders defined in Aime (2006) are also presented. . . . 53

2.3 Distribution of positively selected sites. Distribution of the number of positively selected amino acid sites after correction of p-values with a false discovery method for (a) the data set containing three rust genomes (basidioPAML) and (b) the data set containing the same rust genomes in addition to EST data from *Hemileia vastatrix* (basidioPAML_Hv). Embedded in each histogram is a doughnut chart with the distribution of the positively selected sites across the two main site class pairs defined in this study for the basidioPAML data set (a) and *basidioPAML_Hv* data set (b). *Unique* sites represent amino acids exclusive and identical in all rust species and *Diversifying* sites represent amino acids exclusive but variable in rust species. The site classes are colour coded with the corresponding legend at the bottom of the figure. 56

2.4 Prevalence of positively selected site classes. Pie charts with the distribution of the most prevalent site classes across each positively selected gene for (a) the data set containing only the three rust genomes (basidioPAML) and (b) the data set containing the same rust genomes in addition to EST data from *Hemileia vastatrix* (basidioPAML_Hv). Site classes are colour coded according to the legend in the right. 57

2.5 Enrichment of functional categories among positively selected genes. Bar chart with the fold change comparing the proportion of genes under positive selection and without positive selection in the y-axis, and the several KOG functional categories in the x-axis. For each functional category, fold change values are presented for the basidioPAML and basidioPAML_Hv data sets, according to the legend in the top right corner of the figure. Bars with an asterisk ("*") represent statistically significant results for p-value < 0.1. 59

2.6 Distribution of the number of positively selected amino acid sites after correction of p-values with a false discovery method for (a) the data set containing three rust genomes (basidioPAML) and (b) the data set containing the same rust genomes in addition to EST data from *Hemileia vastatrix* (basidioPAML_Hv). Embedded in each histogram is a doughnut chart with the distribution of the positively selected sites across the three site class pairs defined in this study for (a) the basidioPAML data set and (b) basidioPAML_Hv data set. *Unique* sites represent amino acids exclusive and identical in all rust species, *Diversifying* sites represent amino acids exclusive but variant in rust species and *Conversed* sites represent conserved amino acids across all species. The site classes are colour coded with the corresponding legend on the bottom of the figure. 71

2.7 Pie charts with the distribution of the most prevalent site classes across each positively selected gene for the data set containing (a) only the three rust genomes (basidioPAML) and (b) the data set containing the same rust genomes in addition to EST data from *Hemileia vastatrix* (basidioPAML_Hv). Site classes are colour coded according to the legend in the right. 72

2.8 Bar charts representing the proportion of codon usage for different amino acid residues and data sets between rust and non-rust species. Plots (a) and (c) refer to the data set containing only three rust genomes (basidioPAML), while plots (b) and (d) refer to the data set containing the same three rust genomes in addition to *Hemileia vastatrix* (basidioPAML_Hv). For the serine residue, two plots are shown corresponding to sites strictly conserved (all species possess the same amino acid) and mostly conserved (at least 70% of the species possess the same amino acid). Codons are colour coded according to the legend next to each plot. 73

3.1 Phylogenetic relationships among 29 isolates of *H. vastatrix*. Support values are provided above branches with bootstrap values above 70 and posterior probability above 0.8. For each isolate, information about its geographic origin, pathotype and phylogenetic group is provided. . 93

3.2 Structure plot of the 29 *H. vastatrix*'s isolates with K=3. Vertical bars represent an isolate and the colour proportion for each bar represents the posterior probability of assignment to one of the three clusters. The three groups identified in the phylogenetic tree are outlined above the plot. 94
3.3 Principal component analysis of genomic diversity for 29 isolates of *H. vastatrix*. Isolates are colour coded according to their assignment to the three phylogenetic groups. The three isolates of the *C3* group that revealed a signal of allele sharing with the *C2* group were further differentiated as a fourth *C3Int* group. 95
3.4 Bar plots of the inbreeding coefficient (F_{IT}) for each isolates of *H. vastatrix* from the *C3* (A) and *C2* (B) groups. For each isolate, F_{IT} values were calculated for data sets with (MAF) and without (No MAF) a minor allele frequency filtering. 96
3.5 Summary of the diagnostic SNP scanning for shared alleles between isolates of the *C2* and *C3* phylogenetic groups. The stacked bar plot represents the frequency, while the star point plot represents the percentage, of alleles that isolates shared with the other group. 97
3.6 Venn diagram with the overlap of the SNPs with shared alleles among the three *H. vastatrix* isolates showing admixture signal. 97
3.7 Results from the Index of Association (IA) analysis, using the standardized form (r_d), for the isolates of the *C3* group after removing putative introgressed isolates and an incipient but well supported sub-group. The histogram depicts the distribution of r_d values expected from unlinked loci. The vertical dashed line represents the observed rd value for the data set. 99
3.8 Extended Bayesian skyline plot depicting the population dynamics of the *C3* group of *H. vastatrix* through time. The x-axis is in relative unites of time, and the y-axis corresponds to the effective population size. The dashed black line represents the median estimate of the effective population size, while the solid grey lines delimit the 95% high posterior density. 100

3.9 Effects of the minimum read depth parameter on the four error rates across assemblies. Minimum read depth is provided by the “d” value of the assembly name (*i.e.*, d15 means minimum read depth of 15). . . . 110
3.10 Effects of the clustering threshold parameter on the four error rates across assemblies. Clustering threshold is provided by the “c” value of the assembly name (*i.e.*, c80 means clustering threshold of 0.80). . . 111
3.11 Effects of the maximum shared heterozygosity parameter on the four error rates across assemblies. Maximum shared heterozygosity is provided by the “p” value of the assembly name (*i.e.*, p2 means maximum shared heterozygosity of 2). 112

4.1 Diagram representing TriFusion’s workflow across its three main modules. 126
4.2 Example of the build and testing status of Travis CI for TriFusion. . . 157
4.3 Project quality certification of TriFusion by Codacy. 165
4.4 Conversion and concatenation benchmarks of TriFusion in comparison to other 6 software tools. The “X” symbol represents executions that the corresponding software could not conclude for that data set. “NA” represents particular cases where the conversion of many files is not supported by the corresponding software. The scale of the y-axis is square-root transformed. 166

A.1 Triangular matrix with pairwise F_{ST} comparisons between the phylogenetic groups within *H. vastatrix*. Scatter plots in the upper part of the matrix represent the F_{ST} values for each individual SNP segregating between the given group pair. Histograms in the lower part of the matrix represent the distribution of F_{ST} values for the same segregating SNPs.218
A.2 Allele frequency spectrum for the SNPs of the *C3* phylogenetic group (left) and the *C2* group (right).The vertical dashed line represents the mean of the data set. 219

A.3 Results from the Index of Association analysis, using the standardized form ($\bar{r}_d$), for the isolates of the complete *C3* group (A) and after removing putative introgressed isolates (B). The histogram depicts the distribution of $\bar{r}_d$ values expected from unlinked loci. The vertical dashed line represents the observed $\bar{r}_d$ value for the data set. 219

A.4 Principal component analysis of genomic diversity among the 21 *H. vastatrix* isolates from the *C3* group. Isolates are color coded to differentiate the three introgressed isolates from the remaining members of the *C3* group. 220

A.5 Extended Bayesian skyline plot depicting the population dynamics of the *C3* group of *H. vastatrix* through time with the inferred phylogeny overlapped. The x-axis is in relative units of time, and the y-axis corresponds to the effective population size. The dashed black line represents the median estimate of the effective population size, while the solid grey lines delimit the 95% high posterior density. 220

List of Tables

2.4.1 Description of the data sets assembled, including the number of genes, species alignment columns and the analysis performed. 50

4.8.1 Overview of tests data sets for the *concatenation* benchmarks. 168

4.8.2 Overview of the tests data sets for the *conversion* benchmarks. 168

A.1.1 Details on the genomic and EST data used on the present study. . . . 186

A.1.2 Summary statistics about missing data information and average gene length for the `basidioPAML` data set. 190

A.1.3 Summary statistics about missing data information and average gene length for the `basidioPAML_Hv` data set. 192

A.1.4 Summary statistics about missing data information and average gene length for the `genomic46sp_sparse` data set. 194

A.1.5 Summary statistics about missing data information and average gene length for the `combined67sp_sparse` data set. 197

A.1.6 Summary statistics about missing data information and average gene length for the `genomic46sp_dense` data set. 201

A.1.7 Summary statistics about missing data and average gene length for the `combined67sp_dense` data set. 204

A.1.8 Summary statistics about the prevalence of rusts species for each data set. 208

A.1.9 Detailed table with the results of the branch-site test for each gene contained in the `basidioPAML` and `basidioPAML_Hv` data sets. Accessible online at https://doi.org/10.1371/journal.pone.0143959.s004. . 209

A.1.10 Table containing information regarding the results of the functional annotation, positive selection and evolutionary rate tests for each gene contained in the `basidioPAML` and `basidioPAML_Hv` sets. Accesible online at https://doi.org/10.1371/journal.pone.0143959.s005. . 209

A.1.11 Enrichment analyses overview. Accessible online at https://doi.org/10.1371/journal.pone.0143959.s006. 209
A.2.1 List of the *H. vastatrix* isolates used in this study. 211
A.2.2 Summary of the 11 assemblies of RAD data using **PyRAD** with information on the assembly parameters, error statistics and loci information. 213
A.2.3 Summary statistics of read number and total base pairs for each *H. vastatrix* isolate 214
A.2.4 Summary of the 11 assemblies of RAD data using **PyRAD** with information on the SNP diversity statistics for the *C1+C2* and *C3* groups. . 216
A.2.5 Summary of linkage disequilibrium statistics and results of the significance tests for the complete *C3* data set (`Var1_MM50_maf`), after removing putatively introgressed isolates (`Var1_MM50_NoInt_maf`) and after removing the isolates from the incipient *C3* sub group (`Var1_MM50_NoInt_NoV5_maf`). 217
A.3.1 Benchmark of all operations of **TriFusion**'s Process module for test data sets of diverse compositions. 222

1 Introduction

1.1 The promise of the genomics era

Understanding how genomes evolve and cope with environmental and life-style changes is one of the outstanding questions in evolutionary genomics (Pavey et al., 2012). In this sense, it is difficult to overstate the fundamental impact and potential that the so called Next Generation Sequencing (NGS) revolution had and continues having in biology. This potential was so large that it was compared to the early days of the Polymerase Chain Reaction (PCR) (Metzker, 2010) and was selected by *Nature methods* as the method of the year in 2007 (Schuster, 2008; Ansorge, 2009). In a time where genomic studies were mostly circumscribed to a very few well studied model organisms, the now called High Throughput Sequencing (HTS) technologies paved the way for the application of genome-wide studies to a much broader range of organisms with limited or no previous genomic resources available, thereby allowing the democratization of sequencing (Hudson, 2008; Mardis, 2008a; Pop and Salzberg, 2008; Pavey et al., 2012). As the access to HTS technologies became more affordable, genome-scale sequence data allowed for advanced studies to be carried out on non-model organisms (Mardis, 2008b; Ekblom and Galindo, 2010; Pareek et al., 2011; Grünwald et al., 2016). By shifting genomics from exclusively laboratory-studied model organisms towards studies of natural populations, researchers could start to address novel and important ecological and evolutionary questions that were unrealistic or even inconceivable just a few years ago (Ekblom and Galindo, 2010).

The comparison of entire genomes, either of closely or distantly related species, was perhaps what generated the greatest enthusiasm within the scientific community interested in evolution since the availability of HTS (Hudson, 2008; Pavey et al.,

2012; Chan and Ragan, 2013; Vitti et al., 2013). Upon the first published genome of a eukaryotic species, *Saccharomyces cerevisiae*, in 1996 (Goffeau et al., 1996), sequencing of subsequent genomes accumulated slowly at first with the publication of the *Caenorhabditis elegans* genome in 1998 (C. elegans Consortium, 1998), *Arabidopsis thaliana* in 2000 (The Arabidopsis Genome Initiative, 2000), *Homo sapiens* in 2001 (Venter et al., 2001) and *Mus musculus* in 2002 (Waterston et al., 2002). These were consortia-based projects that required a tremendous amount of resources and relied only on Sanger sequencing, either of bacterial artificial clones or whole genome shotgun libraries (Mardis, 2008a). With the introduction of HTS technologies, there was an astounding growth of genome sequencing projects. By the time this thesis project began, there were 644 eukaryotic genomes sequenced (though not all publicly available), 248 of which were from fungal species (Ellegren, 2014). Since 2018, according to the Genomes OnLine Database (GOLD; https://gold.jgi.doe.gov/; last accessed 23/02/2018), 5 228 genome sequencing projects had been submitted, in varying stages of completion, across 5 049 eukaryotic taxa, 3 124 of which were from fungal species. The access to such an unprecedented number of genomes from multiple species brought the field of evolutionary genomics to a level where inferring evolutionary processes affecting sequence evolution is increasingly performed from a whole-genome perspective rather than at a few loci (Carstens et al., 2012; Chan and Ragan, 2013; Vitti et al., 2013). Indeed, comparative genomics has the potential to directly allow the uncovering of genomic events that led to the origin of a single or entire groups of organisms (Hudson, 2008). Powered by continuous innovation, HTS allowed for nucleotide variation profiling and large scale discovery of genetic markers, which have aided researchers in the pursuit of the genetic basics of ecologically and evolutionary important transitions in organisms (Ekblom and Galindo, 2010; Zhang et al., 2011). Consequently, our understanding of many fundamental biological phenomena has dramatically accelerated over the last decade and HTS will likely continue providing radical insights and change the field of genomics.

Despite the rapid advances, the "*omics*" fields are still emerging or in their infancy when it comes to using whole-genome sequence data to answer biological questions (McCormack et al., 2013). This is particularly true for fields like phylogenomics

and population genomics, where the majority of the underlying theory, methodology and software is still based on developments made in a time where only a very small subset of discrete loci was available to researchers. As a result, they were not designed to take advantage of the full potential of whole-genome data (Kumar et al., 2012; McCormack et al., 2013). Nevertheless, this is a time of a remarkable transition in both sequencing technologies and the methods/software that are used to interpret and understand genome-wide sets of data. In this thesis, two emerging genomic disciplines were leveraged to extend our knowledge on pathogenic fungi at a macro-evolutionary level, using phylogenomics, and micro-evolutionary level, using population genomics.

1.1.1 Phylogenomics

Phylogenomics was originally defined as the use of phylogenetic methods and evolutionary analyses to predict protein functions (Eisen, 1998). However, in the context of molecular systematics, it is also defined as the study of evolutionary relationships of organisms based on comparative analysis of genome-scale data, either in a gene-by-gene or whole-genome basis (Chan and Ragan, 2013). Understanding such relationships is a prerequisite of most evolutionary studies and the basis for inferences in comparative biology (Delsuc et al., 2005; Snel et al., 2005). With the advent of HTS, phylogenomics seemed poised to offer massive benefits to molecular systematics and studies of macro-evolution, as hundreds or even thousands of loci would provide unambiguous resolution and support for all branches in the tree of life (Pyron, 2015). An expected consequence of increasing the quantity of data was that it would dramatically reduce errors due to site sampling and estimation, leading to very high power and confidence values in establishing phylogenetic patterns (Kumar et al., 2012). To some extent, the application of such a wealth of sequence data did resolve or even overturn long-held phylogenetic hypothesis in several taxonomic groups across the tree of life (Jarvis et al., 2014; Misof et al., 2014; Rochette et al., 2014; Zapata et al., 2014; Crawford et al., 2015; Torruella et al., 2015). Paradoxically, unlike the systematics revolution brought by DNA sequencing in the 1990s, the development of genomic data sets did not substantially alter the early molecular

estimates with fewer markers in many cases, providing a clear example that more data is not a panacea (Pyron, 2015; Posada, 2016).

Despite the obvious potential, the application and levering of genome-wide data in phylogenomics was slow to take root in comparison to other fields (McCormack et al., 2013). This lag was attributed to several causes, including the predominant focus on non-model organisms, the need for large numbers of samples (often from divergent taxonomic groups), the transitional state of technology, the predominance of short read data and the complexity of building and executing phylogenomic pipelines (McCormack et al., 2013; Pyron, 2015; Posada, 2016). Moreover, phylogenomic data forced researchers to account for several issues that only became apparent as more data was used. For instance, phylogenetic incongruence between multiple loci became a natural and expected part of the analysis, and researchers needed to discern between "real" biological incongruence and methodological artifacts resulting from a failure of evolutionary models to accommodate complex patterns in the data (Snel et al., 2005; Jeffroy et al., 2006; Betancur et al., 2014). Systematic biases of existing phylogenetic methods and models also started to become evident, often lending strong statistical support to wrong answers (Nishihara et al., 2007; Kumar et al., 2012). These issues, coupled with the challenges of processing large data sets, resulted in the need for strict and, ideally, automated quality control measures of the assembled data just to assure the reliability of the inferences. Many early efforts to facilitate and automate this process were proposed (Philippe et al., 2005; Ciccarelli et al., 2006; Philippe and Blanchette, 2007), but even today it remains open to debate which are the best strategies for the acquisition, manipulation, analysis and interpretation of massive data sets (Posada, 2016). Moreover, many important stages of the data set assembly process were often performed using "in-house custom scripts" that were not easy to access or to use. Consequently, newcomers to the field would come across the stark contrast between an overwhelming abundance of tools for some tasks and near complete absence of support for others, creating a very high entry barrier to the field. This lack of consensus and consolidation was perhaps one of the most important reasons for the slow adoption of phylogenomics.

Despite these challenges, researchers are still pushing the boundaries of the field by developing new techniques that extract informative data from organism's genomes (McCormack et al., 2013) and increasingly sophisticated comparative methods that take full advantage of genome-wide data (Bapteste et al., 2013; Pyron, 2015; Posada, 2016). While the utility of phylogenomics for phylogenetic tree inference is undeniable, the application of genome-scale data remained underappreciated in other related areas, such as the study of genome-wide adaptive molecular evolution (Vitti et al., 2013; Pyron, 2015). Few studies have leveraged the power of genomic data to investigate molecular evolution across the genomes to highlight patterns of adaptation among taxonomically broad groups, even though this pursuit could be performed concurrently with the phylogenetic inference (Pyron, 2015). Until recently, the greatest limitation to this endeavor was the requirement of substantial genomic resources and the limited amount of sequenced genomes available. However, the increasing availability of genome sequencing data in public databases, both in the format of complete or draft genome sequences or as transcriptomes, already makes this a possibility. Moreover, methods that detect selection at the macro-evolutionary level already exist and have been continuously refined (Yang, 2007; Dutheil et al., 2012; Gharib and Robinson-Rechavi, 2013). These typically hinge on comparisons of orthologous sequences among related taxa and identify sequences that are likely to contain functionally relevant differences through the ratio of non-synonymous and synonymous substitutions (d_N/d_S) and/or accelerated rates of evolution (Vitti et al., 2013).

Measuring the d_N/d_S ratio, or ω, provides a direct estimate of whether codons are under selective pressure. For instance, evaluating if ω is significantly higher than 1 constitutes a test of the action of positive selection, while values significantly lower than 1 indicate purifying selection (Anisimova and Liberles, 2007; Vitti et al., 2013). Early approaches used simple counts of d_N and d_S averaged across an entire gene, based on pair-wise sequence comparison, to estimate ω (Miyata and Yasunaga, 1980; Nei and Gojoborit, 1986). However, this averaging leads to a substantial reduction in power to detect positive selection because its signal can vary greatly across the gene and does not occur at a constant rate (Yang and Bielawski, 2000). A molecule can be mostly under strong purifying selection with only a few amino

acid residues being targeted by positive selection, and this may only happen on a particular branch of its evolutionary tree due to an important evolutionary transition. To address these issues, more powerful probabilistic approaches using Maximum Likelihood were developed, allowing for different ω ratios at different sites in the gene and/or lineages in the phylogeny (Yang and Nielsen, 2002; Yang, 2005; Yang and Nielsen, 2008). With these approaches, competing nested models with different statistical distributions of ω are specified, and their parameters estimated in a way that maximizes the likelihood function. Then, a Likelihood Ratio Test (LRT) is used to compare the two competing models and determine which represents a significantly better fit to the data. These comparisons could be made between a null model where ω values above 1 are not allowed (no selection) and an alternative model where the ω ratio can be higher than one (allows selection). If the LRT determines that the model that allows selection is a significant better fit to the data, positive selection is inferred. The most realistic of these methods is the one that allows variation in the selective pressure both across sites and among branches in the phylogeny, also known as the branch-site model (Yang, 2005; Yang, 2011; Gharib and Robinson-Rechavi, 2013; Lu and Guindon, 2013). If positive selection is detected on a particular codon and phylogeny branch, the branch-site model assumes that this signal does not occur elsewhere in the phylogeny and alignment; in other words, the selection signal is exclusive to that branch/site configuration.

The joint application of these methods with genomic information has created an efficient and attractive strategy for finding the causes of many evolutionary and functionally important phenomena driving lineage-specific divergence. Such genome scans for natural selection would forego the traditional requirement of *a priori* knowledge on specific genes of interest and finding and understanding natural selection could transit from hypothesis-testing to hypothesis-generating science (Anisimova and Liberles, 2007; Vitti et al., 2013). Indeed, the power and usefulness of these scans led to their application to a wide range of taxonomic groups to investigate a number of questions, from differential adaptation to disease in humans and chimps (Vamathevan et al., 2008) to salt adaptation in a desert poplar (Ma et al., 2013) and extreme physiological adaptation in turtles (Shaffer et al., 2013). However, the debate is still open on how natural selection shapes genome-wide sequence

evolution, whether it is through gradual changes (gradualism) or punctuated by bursts of lineage-specific changes on a few branches (punctualism) (Orr, 2005); whether changes in protein function are caused by substitutions on a single or a few specific sites, or require substantial changes in its structure. If there is one thing that the history of science has made abundantly clear over the years, is that the truth usually resides somewhere in the middle of opposing schools of thought. Notwithstanding, empirical data will be key to deepen our understanding of genome evolution and natural selection, and this is an area of research where phylogenomics has the potential to provide detailed and crucial information.

1.1.2 Population genomics and RAD sequencing

The term population genomics started to appear in the literature in the 1990s within the context of large scale polymorphism analyses in human populations, and was often viewed as an extension of population genetics with genome-wide sequence data (Stinchcombe and Hoekstra, 2007; Ellegren et al., 2012). The most straightforward contribution of genomics to population genetics was to enormously increase the number of molecular markers, thereby increasing the accuracy and precision of genetic diversity and demographic parameters (Allendorf et al., 2010). In this perspective, population genomics would remove the data generation bottleneck of population genetics studies when addressing the same questions of genome-wide evolutionary and demographic processes affecting population structure (Sedghifar et al., 2016), speciation (Keller et al., 2013) and locus-specific effects that influence adaptation or phenotypes (Croll and McDonald, 2017). However, this reductionist definition hides the fact that population genomics represents a shift from traditional population genetics to an emphasis on both genome-wide and locus-specific effects (Grünwald et al., 2016). The increasing availability of complete genome sequences, coupled with the ability to genotype thousands of single-nucleotide polymorphisms (SNPs) simultaneously, makes it possible to go far beyond traditional population genetics analyses (Wellenreuther and Hansson, 2016). Thus, population genomics represents a new area of research that portends significant conceptual breakthroughs in how we view genetics, evolution, and the emergence of plant pathogens (Grünwald et al., 2016).

The utility of SNPs for population genetics was recognized early on as a valuable tool for revealing the evolutionary history of populations, even before the availability of HTS technologies (Brumfield et al., 2003). Their informative potential and pervasiveness throughout the genome made them ideal markers to investigate the evolutionary history of populations, particularly speciation events, historical demography and population structure (Hickerson et al., 2010; Ellegren, 2014; Leaché and Oaks, 2017). All these events leave signatures in the diversity and allelic frequency of SNPs which, when provided in sufficient number, allow for complex evolutionary scenarios to be unraveled (Pavey et al., 2015; Blanco-Bercial and Bucklin, 2016). For instance, a particularly tricky aspect of fungal pathogen populations is that they are often neither strictly clonal nor sexual, but can have a mixture of both reproduction modes; something that violates some assumptions of classical population genetic methods. If the reproduction mode of a population is not known, a common approach would be to estimate the amount of linkage disequilibrium among SNPs (Milgroom, 1996) via the calculation of the index of association (Agapow and Burt, 2001). With the availability of large SNP data, it became possible to use a standardized form of the index of association which can accommodate mixed reproductive modes by performing a clone correction step (Kamvar et al., 2014). This methodology revealed to be particularly successful and has been extensively applied on fungal populations to understand their often complex reproductive patterns (Goyeau et al., 2007; Alamouti et al., 2011; Gladieux et al., 2011; Kiss et al., 2011; Gladieux et al., 2015; Nieuwenhuis and James, 2016).

Despite their value, it remained a serious technical challenge to retrieve a sufficient number of unlinked SNPs for a given organism, particularly for species without a closely related model organism. For that reason, SNPs remained primarily used in whole-genome linkage and association studies of model species for several years (Lindblad-Toh et al., 2000; Hoskins et al., 2001; Sachidanandam et al., 2001). Even after the introduction of HTS technologies, the majority of the techniques, methodologies and studies were focused on mapping sequencing reads to sequenced genomes of model organisms or closely related taxa (Nielsen et al., 2011). Several efforts were made to bring SNP markers to a widespread use in population genomic studies of non-model organisms (Thomson et al., 2010; Carstens et al., 2012), but it

was not until the development of Reduced Representation Library (RRL) methods, such as Genotype By Sequencing (GBS) (Elshire et al., 2011) or Restriction-site Associated DNA (RAD) (Baird et al., 2008; Baxter et al., 2011; Hohenlohe et al., 2011) that this was achieved. These techniques did for SNP marker development what HTS did for genome sequencing - allowed affordable genome-wide genotyping of many organisms in a single run regardless of the existence of a reference genome or other genomic resources. RAD sequencing (RADseq) in particular, was considered one the most important scientific breakthroughs of the last decade by allowing the development of up to thousands of polymorphic genetic markers in a single experiment (Andrews et al., 2016; Lowry et al., 2017). Consequently, it became the method that made the greatest impact on population genomics and phylogeography to date, and led to an explosion of studies in ecological, evolutionary and conservation genomics (McCormack et al., 2013; Benestan et al., 2015; Gamble et al., 2015; Longo and Bernardi, 2015; Blanco-Bercial and Bucklin, 2016; Hou et al., 2016; Sovic et al., 2016).

RADseq, as previously mentioned, is a reduced representation library technique, which means that it targets only a subset of the genome (Andrews et al., 2016). The term, originally used to describe a particular method (Baird et al., 2008), is now an umbrella for many variations of the original protocol that rely on several key basic steps (Peterson et al., 2012; Wang et al., 2012; Toonen et al., 2013). In the original method, the library preparation process involves the digestion of high molecular weight DNA with one or more restriction enzymes, the subsequent addition of specific sequencing adaptors and, optionally, sequence barcodes that are used to identify individual samples that are sequenced together (Figure 1.1). One of the main advantages of RADseq is that it combines tight control over the fragments resulting from the digestion with high coverage sequencing across many individuals, which makes it one of the most reproducible restriction digest-based methods (McCormack et al., 2013). Depending on the selected restriction enzyme or enzymes, it is also possible to control for the expected number of uncovered markers. For instance, enzymes with shorter recognition sites (*frequent cutters*) will cut the genome more frequently than enzymes with longer recognition sites (*rare cutters*), thereby increasing the number of RAD loci but with the disadvantage of

reducing sequencing depth for the same sequencing effort. Once the libraries are prepared, RADseq uses HTS to generate short read sequence data adjacent to restriction cut sites, thereby generating potentially thousands of SNPs for hundreds of individual samples.

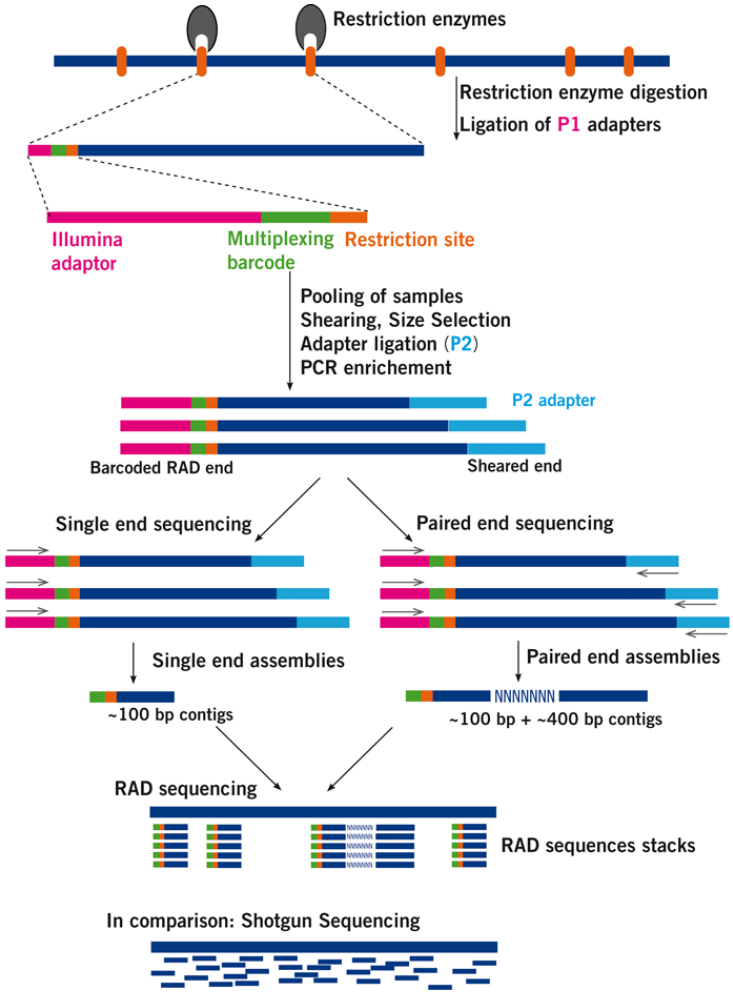

Figure 1.1. Overview of RAD sequencing library preparation and sequencing. Retrieved from https://www.floragenex.com/rad-seq/.

For non-model organisms with scarce genomic resources and/or potentially large and complex genomes, this methodology has several advantages over whole-genome sequencing approaches, such as greater depth of coverage per locus and a more cost-effective approach to sequencing (Andrews et al., 2016). Moreover, the ability to simultaneously screen for polymorphic loci and genotype individual samples greatly mitigated the ascertainment bias of previous SNP analyses, where polymorphic sites were initially screened in a small sample and genotypes where consequently biased towards common and highly variable loci (Brumfield et al., 2003). However, despite all the major advantages that these techniques brought to the field of population

genomics, new and unique sources of error and bias were also introduced, which need to be taken into account. Being a technique based on digestion by restriction enzymes, it is prone to allele dropout and null alleles that occur when a polymorphism resides at a recognition site. Stochastic variation in the PCR stage of HTS is also reported to skew the amplification of one allele more than the other allele, which can lead to downstream genotyping errors (Andrews et al., 2016). Errors related to the preferential amplification based on GC content during PCR are also a well-known phenomenon (Davey et al., 2011). Therefore, genotyping errors are to be expected during the processing of this type of data and researchers need to employ new techniques that minimize the error and maximize the retrieval of informative loci (Mastretta-Yanes et al., 2015). Once these challenges have been addressed and accounted for, population genomics becomes an unparalleled approach for uncovering the evolutionary history and genetic mechanisms underlying the emergence and evolution of populations and species.

1.1.3 The rise of bioinformatics

As sequencing technologies continue to improve with higher sequencing depth and reduction in cost, the computational challenges have grown correspondingly (Egan et al., 2012). Biologists are now faced with the problem of having a hard-drive full of data and begging the question of "now what?". The data throughput that HTS technologies brought resulted in a paradigm shift in the way of how researchers generated and handled such massive amounts of data, to the point where it spurred the development of the field of bioinformatics (Horner et al., 2010; Nekrutenko and Taylor, 2012). Despite emerging back in the 1960s with a different meaning from its current description, bioinformatics has risen to prominence much more recently, as can be seen from the dynamics of bioinformatics-related publications over the last decades (Figure 1.2).

However, the term “bioinformatics” is incredibly interdisciplinary and may refer to many things, such as a field of study, a collection of tools and methodologies or a service (Bartlett et al., 2017). Therefore it is not easy to establish a definition for something that has become an umbrella for a wide range of biological studies

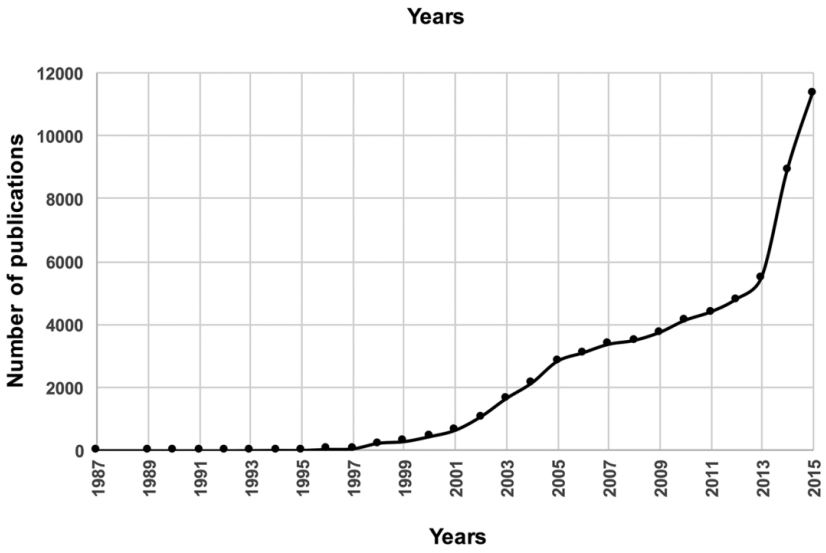

Figure 1.2. Dynamics of bioinformatics related publications over the past four decades with the quoted keyword retrieved from scientific publications in PubMed. Adapted from Atwood et al. (2015).

that targets to develop methodologies and tools to organize, systematize, annotate, visualize, understand and interpret complex data volumes (Abdurakhmonov, 2017). Yet, despite recent attempts of discrediting the field by using the term "research parasites" to describe bioinformaticians and computational biologists (Longo and Drazen, 2016), it is currently a central pillar in life sciences. In fact, most of the current and future challenges seem to lie at the bioinformatics end rather than in the sequence data production (Ekblom and Galindo, 2010; Pavey et al., 2012). This is particularly evident in the "*omics*" fields where the full benefit of HTS technologies cannot be achieved until bioinformatics are able to maximally process and interpret sequence data (Zhang et al., 2011). On a more general note, it has also been proposed that the next modern synthesis in biology will be driven by the merger of mathematical, statistical and computational methods into mainstream biological training, providing every researcher with the essential tools and know-how to navigate the future landscape of biological research (Via et al., 2013; Atwood et al., 2015; Abdurakhmonov, 2017; Markowetz, 2017). Even in today's landscape, it is already essential to have a fundamental understanding of UNIX environments, considered the *lingua franca* of science, and of computer programming languages (e.g., Python, R) to solve context-dependent issues. To paraphrase Hibbett et al. (2013), "the keyboard has joined the pipet and the microscope among the essential resources in the biologist's toolkit".

1.1.4 The genomic revolution for fungal plant pathogens

Fungi are arguably, the most diverse eukaryotes on Earth. Just 20 years ago, the notion that we knew only 6% of the fungal species was treated with skepticism by many biologists. Now, as a consequence of DNA sequence data, it has emerged that this was far from an exaggeration and is actually a conservative estimate of the 1.5-6 million fungal species that are thought to exist (Crous et al., 2015). As part of this diversity, the fungal kingdom includes some of the most devastating pathogens known to humankind (Möller and Stukenbrock, 2017). For instance, in 2007, a routine census of bats hibernating in New York State revealed mass mortalities caused by a fungus growing on their muzzles and wing membranes, later named *Geomyces destructans* (Frick et al., 2010). It was later found that bat numbers across the sampling sites in the United States and Canada had declined by over 70% (Fisher et al., 2012). Another example is the skin infecting amphibian fungus, *Batrachochytrium dendrobatidis*, which was discovered in 1997 and currently infects over 500 species of amphibians worldwide. Its first appearance in the Americas was linked with a wave of population declines that reached over 40% of the populations in some areas of Central America (Fisher et al., 2012). However, plant disease epidemics present the highest impact and have even altered the course of human history multiple times, such as in the late potato blight that led to starvation, economic ruin and the downfall of the English government during the Irish potato famine, and in the twentieth century, when Dutch elm blight and chestnut blight destroyed urban and forest landscapes (Fisher et al., 2012). Nonetheless, one of the most well-known and impressive episodes in history of devastating social and economic consequences brought by a plant disease epidemics was the obliteration of coffee cultivation from Ceylon (Sri Lanka) by Coffee Leaf Rust, which led to a switch in agricultural production from coffee to tea (Cressey, 2013; McCook and Vandermeer, 2015).

Worryingly, the past two decades have seen an increasing number of fungal diseases, some of which have been responsible for the most severe die-offs and extinctions ever witnessed in wild species and presenting the biggest threat to food security worldwide (Fisher et al., 2009; Fisher et al., 2012; Fisher et al., 2016) (Figure

1.3). With the global technological advances the threat of fungal diseases has not abated, but in fact it has been heightened by resource-rich agricultural practices, the globalization of trade and transportation and climate fluctuations (Stukenbrock and McDonald, 2008; Stukenbrock and Bataillon, 2012). Therefore, it has been argued that the nascent fungal infections will cause increasing attrition of biodiversity, with wider implications for human and ecosystem health, unless proper measures are adopted to tighten biosecurity worldwide (Fisher et al., 2012). It has been evident for some time that crop protection from pathogens, namely fungi, is likely to require a significant re-engineering of agro-ecosystems (Stukenbrock and McDonald, 2008). Achieving such a sustainable agriculture is a highly ambitious and elusive goal with many layers of complexity that requires a multidisciplinary approach to establish effective and durable management practices. However, to achieve this, a significant shift in the way we tackle disease management will be required to include ecoevolutionary principles in the design of management programs aimed at minimizing the evolutionary potential of plant pathogens (Zhan et al., 2015). In this aspect, the emergence of genomics seems fit to provide unprecedented advancements in our understanding of fungal pathogenic species and populations (Grünwald et al., 2016).

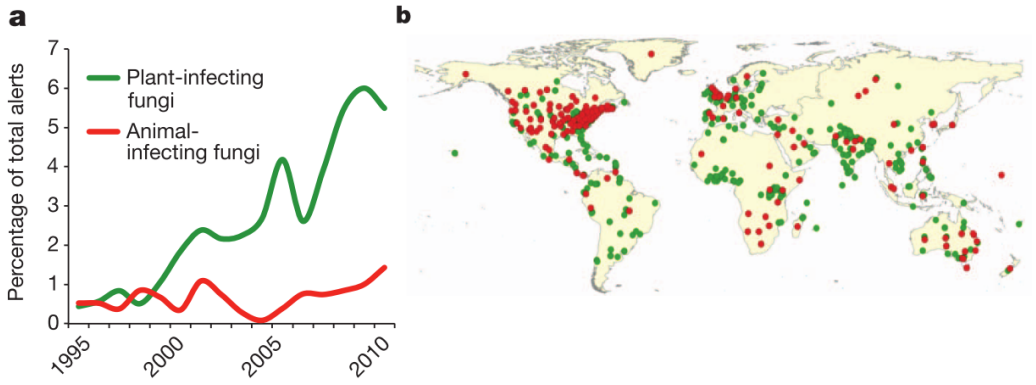

Figure 1.3. Disease alerts in the ProMED database for pathogenic fungi of animals and plants. Retrieved from Fisher et al. (2012).

Among all eukaryotic groups, fungi are the taxonomic group that has been most sequenced so far, owing to their relatively small and compact genomes (Ma et al., 2011). Like all other taxa, fungal genomes accumulated slowly at first but are now being generated at an exponential rate, powered by initiatives like the 1000 Fungal Genomes project launched in 2011, with the aim of sequencing representatives of

roughly 656 families of Fungi (http://1000.fungalgenomes.org). Comparable to other fields of biology, mycology is still transiting into a genome-enabled science, merging HTS and bionformatics to further all aspects of fungal biology. Nevertheless, fields like phylogenetics, systematics and population biology have already been profoundly influenced by genomics. For instance, some of the first phylogenomic studies were focused on fungal taxa, providing a well supported and resolved backbone for several fungal groups (Kuramae et al., 2006; Liu et al., 2009; Fitzpatrick et al., 2006; Robbertse et al., 2006; Martin et al., 2011; Ebersberger et al., 2012). At the macro-evolutionary scale, phylogenomics has been able not only to reconstruct patterns of organismal relationships but also to identify origins of genetic diversity caused by gene families expansion and horizontal gene transfer which may correlate with functional innovations (Slot and Rokas, 2011; Campbell et al., 2012). At the population level, genome-wide association studies based on SNPs have provided the ability to characterize variation at the nucleotide level across the genome and have provided new insights into the mechanisms of adaptation and population divergence in fungi (Neafsey et al., 2010; Ellison et al., 2011; Louis, 2011). Given the profound effect that fungi have on natural and agricultural ecosystems and human health, the synergy of comparative phylogenomic and population genomic analyses have also been successfully used to address the diversity of fungal nutrition modes, metabolic capabilities, host ranges and mechanisms of pathogenesis (Spanu et al., 2010; Floudas et al., 2012; Spanu, 2012; Vogel and Moran, 2013). Pathogen population genomics will find many applications in plant pathology, especially through improving our understanding of pathogen biology. It is also likely to feed information for resistance breeding practices by providing insight into optimal resistance gene deployment strategies (Zhan et al., 2014; Grünwald et al., 2016).

1.2 The rust fungi

Rust fungi are a large, diverse and destructive group of parasites forming a single monophyletic order, the Pucciniales, which comprise approximately one-third of the Basidiomycota phylum (Aime et al., 2006) (Figure 1.4). They are known to infect over 7 000 species of cultivated plants from almost all taxonomic families

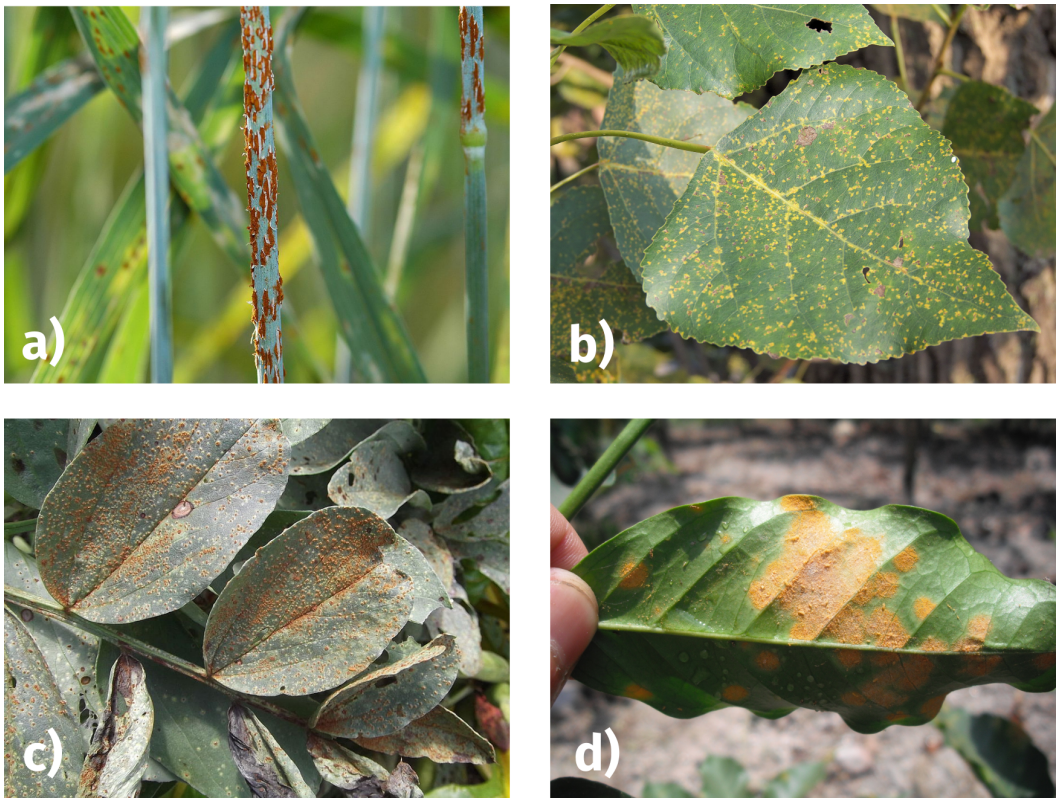

Figure 1.4. Examples of symptomatic (sporulating) infections of rust fungi. **a)** *Puccinia graminis* on wheat, **b)** *Melampsora larici-populina* on poplar, **c)** *Uromyces viciae-fabae* on faba bean and **d)** *Hemileia vastatrix* on coffee.

and the term "rust" refers to the spore coloration that is produced on infected plant organs (Fernandez et al., 2013). In 2012, three species from this group were considered among the top 10 fungal pathogens by the molecular plant pathology journal, including *Puccinia* spp., in the third rank, and *Ustilago maydis* and *Melampsora lini*, in the ninth and tenth ranks, respectively (Dean et al., 2012). The Coffee Leaf Rust pathogen, *Hemileia vastatrix*, is another rust pathogen with a major impact on human society, being considered a natural disaster, but its range is limited to the tropical regions. Despite their importance, the obligate biotrophic life-style and the inability to grow in pure culture or artificial media has hampered research not only on the molecular mechanisms for their pathogenicity, but also on the study of their population biology and evolution. However, HTS technologies have provided a series of recent breakthroughs in our basic understanding of rust fungi biology, particularly on their genomics (Duplessis et al., 2011).

1.2.1 Life-style, life-cycle and infection

In spite of the diversity found among the Pucciniales, all rust species are obligate biotrophs that depend entirely on living host tissues for their growth and reproduction (Fernandez et al., 2013). Biotrophy is a parasitic life-style interaction where the pathogen infects the host without causing the death of the host cells and tissues, and this is a pervasive and polyphyletic trait that has evolved independently in many fungi and oomycetes (Spanu, 2012). These types of pathogens are also unable to survive without a living host and need to maintain host cells alive in order to complete their life-cycle. This contrasts with necrotrophy, where the pathogen infection leads to the death of the host's cells from which it feeds, and saprotrophy, where the microbe feeds of already dead or decayed host tissue. While it is apparently easy to categorize these life-styles, in fact there is a continuum without hard boundaries. Even archetypal necrotrophs such as *Botrytis cinerea* have an initial biotrophic phase in which they colonize host tissues asymptomatically before unleashing large scale host death (Spanu, 2012). Interestingly, the appreciation that all fungal plant pathogens have an initial biotrophic infection phase has been put forward as evidence that biotrophy is the ancestral pathogenic state and an adaptive valley that pathogens are prone to falling into (Spanu, 2012).

In terms of the life cycle of rust fungi, its complexity varies considerably between different species and ecological conditions (Staples, 2000). Macrocyclic rusts are the most complex examples as they go through five spore stages during their life cycle that lead to the production of five types of spores: basidiospores, spermatia, aeciospores, urediniospores and teliospores. However, life-cycles of rust fungi are plastic and they may have reduced life-cycles that lack one or more of these spore stages (Fernandez et al., 2013; Aime et al., 2017). Those that lack the uredinial and aecial stages (and in some instances the spermogonial stage as well) are microcyclic, while those that lack only the uredinial stage are called demicyclic. The final variation, hemicyclic, is used to describe species without known aecial or spermogonial stages. In addition to this classification, rust species can be heteroecious when they require more than two alternate hosts species for the completion of their life-cycle. On the other hand, autoecious rusts can complete the entire life-cycle in a single host

species. Therefore, individual species can have a combination of these life-cycle traits, such as the macrocyclic heteroecious rusts *Puccinia* spp. and *Melampsora* spp., or the hemicyclic autoecious *Hemileia vastatrix* (Aime et al., 2017).

Regardless of the life-cycle, the basic unit of infection and dissemination of rust fungi is the spore, a dykaryotic urediniospore, that represents the asexual cycle. In nature, the dissemination process of the urediniospores is deeply dependent on the wind and rain. Events such as winds and storms are responsible for long distance dissemination, while rain, mainly rain splashes, is responsible for the dispersion at short distances (Staples, 2000). Once the urediniospore adheres to the surface of a susceptible host, a series of host-surface recognition steps lead to the spore germination and appressorium formation over stomata, or more rarely over the epidermis, before penetration and inter- and intracellular colonization. After host penetration, the hallmark of biotrophy is the formation of a specialized structure called haustorium, which is mainly dedicated to nutrient uptake (Voegele and Mendgen, 2011). Recent findings also suggest that haustoria are able to suppress the host defense response and reprogram host immunity and metabolism by secreting effector molecules (Duplessis et al., 2011; Spanu, 2012; Fernandez et al., 2013). When the haustorium is functional, nutrients start to be transported to the epiphytic hyphae to feed the colony. Some days after the infection, uredinia (rust pustules of uredosporic sori) rupture from the host epidermis or stomata, and abundant asexual urediniospores are released to propagate the infection. Species infecting seasonal hosts may develop dikaryotic teliospores during the host's growth season, which will serve as overwinter structures. In perenial hosts, polycyclic uredinial infection cycles can persist throughout the year. In general, rusts reproduction is a complex process that is essentially reliant on multiple, polycyclic rounds of asexual propagation as well as a sexual cycle once a year, when it exists, with the potential for generating genetic variation through recombination.

1.2.2 Pathogenicity and host interaction

Plants in nature are generally resistant to most pathogens, with the ability to recognize potential invading organisms and with elaborate defense mechanisms to ward

off pathogen attack (Aguileta et al., 2009; Schie and Takken, 2014; Balloux and Dorp, 2017). Therefore, a high degree of adaptation is required by the pathogen to be able to colonize a host and overcome its defenses. In this sense, pathogens need to continuously develop strategies to compromise plant resistance mechanisms in a perpetual interaction that is often called the "Red Queen" dynamics, as a metaphor for an evolutionary arms race (Clay and Kover, 1996). This interaction forces the host and pathogen to co-evolve, resulting in a high degree of host specialization, a phenomenon that is particularly widespread among biotrophic pathogens that have a long-lasting interaction with the host's living cells (Schie and Takken, 2014). In response to pathogen attacks, plants have evolved at least two lines of active defenses (Jones and Dangl, 2006; De Wit et al., 2009; Boyd et al., 2013). The first line provides basal defense against all potential pathogens and is based on the recognition of conserved pathogen-associated molecular patterns (PAMPs), by so-called pattern recognition receptors (PRRs) that activate PAMP-triggered immunity (PTI). Successful pathogens deliver effector proteins to suppress PTI. The second layer of defense is activated when a given pathogen-derived effector, named Avr protein, is 'specifically recognized' by plant receptor proteins encoded by *R* genes, resulting in effector-triggered immunity (ETI). This *R*/*Avr* recognition has been termed gene-for-gene resistance (Flor, 1956). Whereas the first response confers resistance based on a broad recognition of evolutionary conserved molecules, the second response is mostly species or race specific (Schie and Takken, 2014).

Rust fungi are prime examples of host and pathogen co-evolution and their interaction has been studied for decades. The cumulative results of these studies led to the first genetic statements of the particular plant-pathogen interaction between the rust fungus *Melampsora lini* and flax in the historical landmark publication by Harold H. Flor (Flor, 1956). Half a century after his work, numerous studies have so far mostly supported the gene-for-gene model as the mechanism governing plant-rust interactions in general (Dodds, 2004; Fernandez et al., 2013). Moreover, it was also demonstrated that the genes encoding the Avr proteins are under strong directional selection with high levels of amino acid polymorphism, which is consistent with the arms race that hosts and pathogens are continuously engaging (Schie and Takken, 2014). An example is the *AvrP4* across multiple species of the Melampsora genus

(Van der Merwe et al., 2009). This selective pressure has resulted in the emergence of both rust races with different virulence profiles, and resistant host varieties with different sets of resistance genes. For instance, around 50 rust resistance genes have been identified in wheat, more than nine resistance factors in coffee and eight qualitative resistances factors in poplar (Fernandez et al., 2013). However, the genetic determinants of both *Avr* and *R* genes have been identified only in a very limited subset of species and it is important to continue to verify to which extent the gene-for-gene model holds in plant-rust fungi interactions. This is particularly urgent since variations from the classic gene-for-gene model have been reported, with examples of recessive resistance or indirect interactions between the *R* and *Avr* gene products (Jones and Dangl, 2006). With the recent availability of several draft genome sequences for several rust fungi, and the constant development of tools for genetic transformation of rust fungi, it should be possible to identify new putative avirulence genes in the near future.

1.2.3 Genomic features

The sequencing of the first rust genomes was launched in 2006 for the wheat stem rust, *P. graminis* f. sp. *tritici* and in 2007 for the poplar leaf rust, *M. larici-populina*. The first surprising finding from these projects was that unlike the small and compact genomes of most fungi, the size and complexity of the rust fungi genomes was staggeringly high (Duplessis et al., 2011). In fact, it was later found that the Pucciniales have the largest genomes of the fungal kingdom, with an average genome size of 305.5Mbp and reaching the maximum value of 893.2Mbp in *Gymnosporangium confusum* (Tavares et al., 2014). The number of protein coding genes for both *P. graminis* and *M. larici-populina* genomes were similarly high compared to other sequenced basidiomycetes, with ~17k and ~16k genes, respectively, but with no evidence of genome-scale duplication. However, the increase in the genome size is mostly attributed to a proliferation of repetitive DNA and transposable elements (TE), similarly to other fungal biotrophs, with proportions reaching as high as 75% of the genome size (Cristancho et al., 2014). Interestingly, there is no evidence of compartmentalization of repetitive DNA in these genomes (Spanu, 2012).

Manual annotation of rust genomes revealed expansions and contractions of gene families with several biological functions that partly explain the increased genome size (Duplessis et al., 2011; Cantu et al., 2013; Cristancho et al., 2014; Nemri et al., 2014). These changes include the expansion of lineage-specific gene families, a large repertoire of effector-like small secreted proteins with unknown function, absence of sucrose transporters as well as genes involved in inorganic nitrogen and sulphur uptake and assimilation, and a reduced number of carbohydrate active enzymes. Among these, the small secreted proteins correspond to many virulence effectors as well as avirulence factors of other plant pathogens, making them prime targets to study rust pathogenicity (Duplessis et al., 2011). These genome features are thought to be related to their obligate biotrophic life-style, some of which are mirrored by the genomes of other biotrophic organisms, indicating convergent adaptation to biotrophy (Duplessis et al., 2011; Spanu, 2012). Strikingly, less than 50% of the predicted proteins showed significant homology to GenBank's protein database, a feature that seems to be singular to the rusts. Moreover, comparisons between the *P. graminis* f. sp. *tritici*, *M. larici-populina* and *P. striiformis* f. sp. *tritici* revealed proportions of homologous genes between 57 and 69% (Fernandez et al., 2013). This suggests that while a substantial portion of the genome is shared among the Pucciniales, a significant part of each genome seems to be highly species specific and mostly likely related to adaptations to their hosts.

Even though the discovery of these genome-scale changes significantly advanced our knowledge on rust fungi, most of these efforts were focused on finding new or exclusive genomic features that were unique to each rust genome. Considerably less explored remains the role of adaptive genetic variation on shared and conserved genes among rust fungi and other basidiomycetes, but the increasing availability of genomic resources in the Pucciniales group has opened an inviting portal to investigation.

1.2.4 *Hemileia vastatrix* - Coffee Leaf Rust

Hemileia vastatrix, the causal agent of the devastating Coffee Leaf Rust (CLR), belongs to a phylogenetically basal genus of the Pucciniales order with at least 42

species found in tropical and sub-tropical regions of Africa and Asia (Ritschel, 2005). *H. vastatrix* is the type species of the genus and a hemicyclic fungus that produces urediniospores, teliospores and basidiospores, though the asexual urediniospores are the only functional propagules (Talhinhas et al., 2017). The pathogen infects several species of the *Coffea* genus, but only *C. arabica* and *C. canephora* are economically relevant worldwide (McCook and Vandermeer, 2015). Of these two hosts, *C. arabica* is the most economically valuable species and the most susceptible to *H. vastatrix* attacks. CLR disease causes premature defoliation of the plant, leading to yield losses of up to 30% on susceptible coffee plant varieties, if no control measures are applied (Hindorf and Omondi, 2010). Host plants are rarely killed from CLR infections but their yield is severely affected in subsequent years because the vegetative development is hindered. Therefore, *H. vastatrix* is considered the most important pathogen of coffee crops (*Coffea* spp.) worldwide and CLR is amongst the most serious crop diseases in history (Talhinhas et al., 2017).

The disease was first recorded in 1861 in East Africa on wild *Coffea*, and the first disease epidemic occurred soon after when it quickly devastated coffee plantations from Sri Lanka, eventually eradicating its cultivation from the country (Webb, 2002). Such an aggressive initial outbreak was regarded as the outcome of the intense process of coffee domestication from a very narrow gene pool that was most probably free of rust. Since then, and in just 150 years, CLR spread to most coffee-growing countries worldwide, presumably carried by wind currents (McCook and Vandermeer, 2015). In 2008, CLR has attracted even more notoriety due to the successive occurrence of a series of cluster outbreaks across the Americas, collectively described as the Big Rust (Avelino et al., 2015; McCook and Vandermeer, 2015; Zambolim, 2016). The exacerbation of the epidemics has been associated to recent meteorological anomalies linked to climate change that extend both the temporal and spatial range of the disease, resulting in the attribution of natural disaster status to CLR in the tropics (Avelino et al., 2015).

Breeding for coffee resistance is considered to be the best long term solution to control CLR, both environmentally and economically (Silva et al., 2006; McCook and Vandermeer, 2015). The identification and characterization of a naturally resistant

hybrid between *C. arabica* and *C. canephora*, named "Timor's Hybrid" (HDT) has been the basis for the majority of the breeding programs that released rust-resistant cultivars globally. These have been pioneered and propelled by several initiatives and research programs at the Coffee Rusts Research Center (CIFC) in Portugal. Populations of HDT were first retrieved from the island of Timor in 1927, exhibiting resistance to all rust races know at that time. CIFC used these plants to develop breeding programmes aiming at transferring resistance from HDT into the main Arabica cultivars from the coffee-growing countries. Several well succeeded hybrids synthesized at CICF, that combine the resistance of HDT and the good agronomic traits of commercial varieties, were then deployed in the field and currently represent the majority of the cultivated rust-resistant varieties. Despite the success of these resistant varieties, some of which retained complete resistance for over 30 years, their introduction in the field have inevitably resulted in the loss of resistance due to adaptation of the pathogen (Talhinhas et al., 2017).

Coffee-rust interaction follows Flor's gene-for-gene model, and resistance is conditioned by at least 9 resistance genes (Noronha-Wagner and Bettencourt, 1967). Rust resistance in HDT populations is conferred by four *C. canephora*-derived genes (S_H6-9) and others not yet identified, in addition to four *C. arabica*-derived resistance genes (S_H1, S_H2, S_H4, S_H5). These together with the single *C. liberica*-derived gene (S_H3), make up the full spectrum of coffee differential genotypes. Partial and non-specific polygenic resistance has also been reported in *C. canephora* and some *C. arabica* varieties, suggesting that in addition to the resistance genes, other major and minor genes might influence coffee-rust interactions (Silva et al., 2006). It is worth noting, however, that these are resistance genes inferred from classical genetic experiments but have not yet been isolated (Talhinhas et al., 2017). Likewise, *H. vastatrix* race genotypes are inferred by Flor's gene-to-gene theory as comprising distinct and unique combinations of virulence genes, ranging from v_1 to v_9 in isolates derived from *C. arabica* and tetraploid interspecific hybrids, while those of the races that attack diploid coffee species are not known. This nomenclature allows for direct correlation between host resistance and pathogen virulence, since virulence genes v_1 *to* v_5 can be traced back to Arabica-type hosts, whilst genes v_6 to v_9 reflect more recent *C. canephora*-type resistance heritage. However, virulence

profiling of *H. vastatrix* is only possible insofar as the available collection of coffee differential genotypes. As presumed from the continuous dynamic of host-pathogen co-evolutionary arms race, short-term selection on pathogen strains will promote the emergence of new virulence factors. In fact, the constant exertion of selective pressure on the pathogen caused by the introduction of coffee resistant varieties in the field has led to the emergence of more than 50 rust races, or pathotypes (McCook and Vandermeer, 2015; Talhinhas et al., 2017). Moreover, hyper virulent *H. vastatrix* isolates able to infect coffee genotypes previously resistant to all known rust races have already been identified in India (Prakash et al., 2014).

The seriousness of CLR epidemics has triggered emergency actions across several coffee producing countries and investigation of the pathogen's biology gained a considerable momentum, particularly after the beginning of this thesis. A partial and hybrid genome assembly of eight distinct *H. vastatrix* isolates was carried out, but achieving a complete draft genome was difficult due to a remarkably high proportion of repetitive content and unexpectedly high genome size (Cristancho et al., 2014). In addition, transcriptomic data from different life stages of the fungus have been made available (Talhinhas et al., 2014). However, research on the evolutionary history of *H. vastatrix* remains considerably less explored, particularly at a global scale. Several efforts aimed at uncovering suitable molecular markers for characterizing *H. vastatrix*'s genetic variation have been conducted, but RAPDs and AFLPs remain the only informative markers reported so far (Talhinhas et al., 2017). Using these markers, population genetic studies on *H. vastatrix* have predominantly focused on geographically restricted populations, such as those in Brazil (Nunes et al., 2009; Maia et al., 2013; Cabral et al., 2016) or Colombia (Rozo et al., 2012), with the exception of (Gouveia et al., 2005), who included isolates from Asia, Africa and America. An overarching consensus across these studies is that there is no evidence of population structure with respect to race/pathotype, host or geographic region. To the extent of the available sampling, *H. vastatrix* is considered to be a large unstructured species, whose spores are capable of traveling very long distances by wind or mediated by human activities. Moreover, the link between the high phenotypic variability of *H. vastatrix* pathotypes and its genetic diversity has been hard to establish. Estimates of genetic variation have ranged from low (Rozo et al.,

2012; Cabral et al., 2016), to moderate (Gouveia et al., 2005; Maia et al., 2013) and high (Nunes et al., 2009). Similar discrepancies have been found regarding the sexuality of *H. vastatrix*. *H. vastatrix* has been considered an asexual pathogen based on the fact that the sexual phase of its life-cycle was not identified, and that the asexual urediniospores are the only known functional propagules. However, meiosis has been recently suggested to occur within the urediniospores in a supposedly hidden sexual reproductive cycle (Carvalho et al., 2011). Whether this translates into recombination effectively occurring in natural populations, remains an open question. Studies have reported mixed results with some failing do detect recombination and supporting the asexual status of *H. vastatrix* (Gouveia et al., 2005; Rozo et al., 2012), while others have found evidence of recombination in some specific sub-groups (Maia et al., 2013; Cabral et al., 2016). These stark differences in results are likely to be a reflection of the fragmented sampling strategies and spore collection techniques but they do reveal the state of confusion and uncertainty that underlies our understanding of *H. vastatrix* evolutionary history and potential.

1.3 Objectives

This thesis intended to harness the informative power of genome-scale data to improve our understanding on fungal evolution, integrating both macro- and micro-evolutionary perspectives. At the macro-evolutionary scale, the focus was on the application of a phylogenomics approach to reconstruct the evolutionary relationships among several basidiomycetes, with particular focus on the rust fungi, and then study the patterns of genome-wide positive selection at the origin of the Pucciniales. At the micro-evolutionary scale, the focus was on the application of population genomic methodologies to understand the evolutionary history of a particular rust fungus, *Hemileia vastatrix*.

Specifically, this project consisted of three main tasks/objectives:

1. Evaluate the role of positive selection acting on the genomic landscape shared by most basidiomycetes on the root of the Pucciniales. Genes detected as

being under positive selection could be prime candidates to explain the life-style transition to biotrophy and pathogenicity that was seen in the Pucciniales.

2. Use RAD sequencing to investigate the evolutionary history of the rust fungus, *H. vastatrix*, and assess at a genome-wide scale its evolutionary potential, genetic structure and variation, with the final goal of providing information for adjusted and enhanced disease management strategies.

3. Given the tremendous amount of software developed to complete the previous objectives, and considering the contemporary lack of software tools to perform similar tasks, this task aimed at developing a feature rich application designed to gather, process and visualize large scale data matrices for phylogenomics and population genomics.

This thesis document comprises five parts. Chapter 1, contains a general introduction that contextualizes the overall themes covered in this thesis. Chapter 2, contains the results of the phylogenomic approach that addresses the first objective and were published in PLoS One. It also includes additional results and discussion of positive selection affecting conserved amino acid sites that were not included in the publication. Chapter 3, contains the results of the population genomics approach, also published in Molecular Plant Pathology. Chapter 4, contains the outcome of the TriFusion software development (manuscript submitted to a peer-review journal) and a general description of the software. Finally, chapter 5 contains the conclusion and final remarks of the thesis project.

1.4 References

Abdurakhmonov, I. (2017). „Bioinformatics: Basics, Development, and Future“. *Intech*, p. 450.

Agapow, P. and A Burt (2001). „Indices of multilocus linkage disequilibrium“. *Molecular Ecology Notes* 1.1, pp. 101–102.

Aguileta, G., M. E. Hood, G. Refrégier, and T. Giraud (2009). „Genome evolution in plant pathogenic and symbiotic fungi“. *Advances in Botanical Research* 49, pp. 151–193.

Aime, M. C., P. B. Matheny, D. A. Henk, E. M. Frieders, R. H. Nilsson, M. Piepenbring, D. J. McLaughlin, L. J. Szabo, D. Begerow, J. P. Sampaio, et al. (2006). „An overview of the higher level classification of Pucciniomycotina based on combined analyses of nuclear large and small subunit rDNA sequences." *Mycologia* 98.6, pp. 896–905.

Aime, M. C., A. R. McTaggart, S. J. Mondo, and S. Duplessis (2017). „Phylogenetics and Phylogenomics of Rust Fungi". In: *Advances in Genetics*. Vol. 100, pp. 267–307.

Alamouti, S. M., V. Wang, S. Diguistini, D. L. Six, J. Bohlmann, R. C. Hamelin, N. Feau, and C. Breuil (2011). „Gene genealogies reveal cryptic species and host preferences for the pine fungal pathogen Grosmannia clavigera." *Molecular ecology* 20.12, pp. 2581–2602.

Allendorf, F. W., P. A. Hohenlohe, and G. Luikart (2010). „Genomics and the future of conservation genetics." *Nature Reviews Genetics* 11, pp. 697–709.

Andrews, K. R., J. M. Good, M. R. Miller, G. Luikart, and P. A. Hohenlohe (2016). „Harnessing the power of RADseq for ecological and evolutionary genomics". *Nature Review Genetics* 17, pp. 81–92.

Anisimova, M and D. A. Liberles (2007). „The quest for natural selection in the age of comparative genomics." *Heredity* 99, pp. 567–579.

Ansorge, W. J. (2009). „Next-generation DNA sequencing techniques." *New Biotechnology* 25, pp. 195–203.

Atwood, T. K., E. Bongcam-Rudloff, M. E. Brazas, M. Corpas, P. Gaudet, F. Lewitter, N. Mulder, P. M. Palagi, M. V. Schneider, C. W. van Gelder, et al. (2015). „GOBLET: The Global Organisation for Bioinformatics Learning, Education and Training". *PLoS Computational Biology* 11, pp. 1–10.

Avelino, J., M. Cristancho, S. Georgiou, P. Imbach, L. Aguilar, G. Bornemann, P. Läderach, F. Anzueto, A. J. Hruska, and C. Morales (2015). „The coffee rust crises in Colombia and Central America (2008–2013): impacts, plausible causes and proposed solutions". *Food Security* 7.2, pp. 303–321.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko, and E. A. Johnson (2008). „Rapid SNP discovery and genetic mapping using sequenced RAD markers". *PLoS ONE* 3.10, pp. 1–7.

Balloux, F. and L. van Dorp (2017). „What are pathogens, and what have they done to and for us?" *BMC Biology* 15.1, p. 91.

Bapteste, E., L. van Iersel, A. Janke, S. Kelchner, S. Kelk, J. O. McInerney, D. A. Morrison, L. Nakhleh, M. Steel, L. Stougie, et al. (2013). „Networks: expanding evolutionary thinking". *Trends in Genetics* 29.8, pp. 8–10.

Bartlett, A., B. Penders, and J. Lewis (2017). „Bioinformatics: indispensable, yet hidden in plain sight?" *BMC Bioinformatics* 18.1, p. 311.

Baxter, S. W., J. W. Davey, J. S. Johnston, A. M. Shelton, D. G. Heckel, C. D. Jiggins, and M. L. Blaxter (2011). „Linkage Mapping and Comparative Genomics Using Next-Generation RAD Sequencing of a Non-Model Organism". *PLoS ONE* 6.4, e19315.

Benestan, L., T. Gosselin, C. Perrier, B. Sainte-Marie, R. Rochette, and L. Bernatchez (2015). „RAD-genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species; the American lobster (Homarus americanus)". *Molecular Ecology* 24.13, pp. 3299–3315.

Betancur, R., G. J. P. Naylor, and G. Ortí (2014). „Conserved genes, sampling error, and phylogenomic inference.“ *Systematic Biology* 63.2, pp. 257–62.

Blanco-Bercial, L. and A. Bucklin (2016). „New view of population genetics of zooplankton: RAD-seq analysis reveals population structure of the North Atlantic planktonic copepod Centropages typicus“. *Molecular Ecology* 25.7, pp. 1566–1580.

Boyd, I. L., P. H. Freer-Smith, C. A. Gilligan, and H. C. J. Godfray (2013). „The Consequence of Tree Pests and Diseases for Ecosystem Services“. *Science* 342.6160, pp. 1235773–1235773.

Brumfield, R. T., P. Beerli, D. A. Nickerson, and S. V. Edwards (2003). „The utility of single nucleotide polymorphisms in inferences of population history“. *Trends in Ecology and Evolution* 18.5, pp. 249–256.

C. elegans Consortium, T. (1998). „Genome sequence of the nematode C. elegans: A platform for investigating biology“. *Science* 282.5396, pp. 2012–2018.

Cabral, P. G. C., E. Maciel-Zambolim, S. A. S. Oliveira, E. T. Caixeta, and L. Zambolim (2016). „Genetic diversity and structure of Hemileia vastatrix populations on Coffea spp.“ *Plant Pathology* 65.2, pp. 196–204.

Campbell, M. A., A. Rokas, and J. C. Slot (2012). „Horizontal transfer and death of a fungal secondary metabolic gene cluster“. *Genome Biology and Evolution* 4.3, pp. 289–293.

Cantu, D., V. Segovia, D. MacLean, R. Bayles, X. Chen, S. Kamoun, J. Dubcovsky, D. G. O. Saunders, and C. Uauy (2013). „Genome analyses of the wheat yellow (stripe) rust pathogen Puccinia striiformis f. sp. tritici reveal polymorphic and haustorial expressed secreted proteins as candidate effectors.“ *BMC Genomics* 14, p. 270.

Carstens, B., A. R. Lemmon, and E. M. Lemmon (2012). „The promises and pitfalls of next-generation sequencing data in phylogeography“. *Systematic Biology* 61.5, pp. 713–715.

Carvalho, C. R., R. C. Fernandes, G. M. A. Carvalho, R. W. Barreto, and H. C. Evans (2011). „Cryptosexuality and the Genetic Diversity Paradox in Coffee Rust, Hemileia vastatrix“. *PLoS ONE* 6.11, e26387.

Chan, C. X. and M. A. Ragan (2013). „Next-generation phylogenomics.“ *Biology Direct* 8.3, p. 3.

Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork (2006). „Toward automatic reconstruction of a highly resolved tree of life.“ *Science* 311.5765, pp. 1283–1287.

Clay, K. and P. X. Kover (1996). „the Red Queen Hypothesis and Plant/Pathogen Interactions“. *Annual Review of Phytopathology* 34.1, pp. 29–50.

Crawford, N. G., J. F. Parham, A. B. Sellas, B. C. Faircloth, T. C. Glenn, T. J. Papenfuss, J. B. Henderson, M. H. Hansen, and W. B. Simison (2015). „A phylogenomic analysis of turtles“. *Molecular Phylogenetics and Evolution* 83, pp. 250–257.

Cressey, D. (2013). „Coffee rust regains foothold Researchers marshal technology in bid to thwart fungal outbreak in Central America“. *Nature* 493.7434, p. 587.

Cristancho, M. A., D. O. Botero-Rozo, W. Giraldo, J. Tabima, D. M. Riaño-Pachón, C. Escobar, Y. Rozo, L. F. Rivera, A. Durán, S. Restrepo, et al. (2014). „Annotation of a hybrid partial genome of the coffee rust (Hemileia vastatrix) contributes to the gene repertoire catalog of the Pucciniales.“ *Frontiers in Plant Science* 5, p. 594.

Croll, D. and B. A. McDonald (2017). „The genetic basis of local adaptation for pathogenic fungi in agricultural ecosystems“. *Molecular Ecology* 26.7, pp. 2027–2040.

Crous, P. W., D. L. Hawksworth, and M. J. Wingfield (2015). „Identifying and Naming Plant-Pathogenic Fungi: Past, Present, and Future“. *Annual Review of Phytopathology* 53.1, pp. 247–267.

Davey, J, P Hohenlohe, P Etter, J Boone, J Catchen, and M Blaxter (2011). „Genome-wide genetic marker discovery and genotyping using next-generation sequencing“. *Nature Reviews Genetics* 12.7, pp. 499–510.

De Wit, P. J.G. M., R. Mehrabi, H. A. Van Den Burg, and I. Stergiopoulos (2009). „Fungal effector proteins: Past, present and future“. *Molecular Plant Pathology* 10.6, pp. 735–747.

Dean, R., J. A. Van Kan, Z. A. Pretorius, K. E. Hammond-Kosack, A. Di Pietro, P. D. Spanu, J. J. Rudd, M. Dickman, R. Kahmann, J. Ellis, et al. (2012). „The Top 10 fungal pathogens in molecular plant pathology“. *Molecular Plant Pathology* 13.4, pp. 414–430.

Delsuc, F., H. Brinkmann, and H. Philippe (2005). „Phylogenomics and the reconstruction of the tree of life“. *Nature Reviews Genetics* 6.5, pp. 361–375.

Dodds, P. N. (2004). „The Melampsora lini AvrL567 Avirulence Genes Are Expressed in Haustoria and Their Products Are Recognized inside Plant Cells“. *the Plant Cell Online* 16.3, pp. 755–768.

Duplessis, S., C. A. Cuomo, Y.-C. Lin, A. Aerts, E. Tisserant, C. Veneault-Fourrey, D. L. Joly, S. Hacquard, J. Amselem, B. L. Cantarel, et al. (2011). „Obligate biotrophy features unraveled by the genomic analysis of rust fungi.“ *Proceedings of the National Academy of Sciences of the United States of America* 108.22, pp. 9166–9171.

Dutheil, J. Y., N. Galtier, J. Romiguier, E. J. Douzery, V. Ranwez, and B. Boussau (2012). „Efficient selection of branch-specific models of sequence evolution.“ *Molecular Biology and Evolution* 29.7, pp. 1861–1874.

Ebersberger, I., R. de Matos Simoes, A. Kupczok, M. Gube, E. Kothe, K. Voigt, and A. von Haeseler (2012). „A consistent phylogenetic backbone for the fungi.“ *Molecular Biology and Evolution* 29.5, pp. 1319–1334.

Egan, a. N., J. Schlueter, and D. M. Spooner (2012). „Applications of next-generation sequencing in plant biology“. *American Journal of Botany* 99.2.

Eisen, J. A. (1998). „Phylogenomics : Improving Functional Predictions for Uncharacterized Genes by Evolutionary“. *Genome Research* 8, pp. 163–167.

Ekblom, R and J Galindo (2010). „Applications of next generation sequencing in molecular ecology of non-model organisms.“ *Heredity* 107, pp. 1–15.

Ellegren, H. (2014). „Genome sequencing and population genomics in non-model organisms“. *Trends in Ecology and Evolution* 29.1, pp. 51–63.

Ellegren, H., L. Smeds, R. Burri, P. I. Olason, N. Backström, T. Kawakami, A. Künstner, H. Mäkinen, K. Nadachowska-Brzyska, A. Qvarnström, et al. (2012). „The genomic landscape of species divergence in Ficedula flycatchers.“ *Nature* 491.7426, pp. 756–760.

Ellison, C. E., C. Hall, D. Kowbel, J. Welch, R. B. Brem, N. L. Glass, and J. W. Taylor (2011). „Population genomics and local adaptation in wild isolates of a model microbial eukaryote.“ *Proceedings of the National Academy of Sciences of the United States of America* 108.7, pp. 2831–2836.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. a. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell (2011). „A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species." *PloS ONE* 6.5, e19379.

Fernandez, D., P. Talhinhas, and S. Duplessis (2013). „Rust Fungi: Achievements and Future Challenges on Genomics and Host–Parasite Interactions". In: *Agricultural Applications*. Vol. 11, pp. 315–341.

Fisher, M. C., T. W. J. Garner, and S. F. Walker (2009). „Global emergence of Batrachochytrium dendrobatidis and amphibian chytridiomycosis in space, time, and host." *Annual Review of Microbiology* 63, pp. 291–310.

Fisher, M. C., D. a. Henk, C. J. Briggs, J. S. Brownstein, L. C. Madoff, S. L. McCraw, and S. J. Gurr (2012). „Emerging fungal threats to animal, plant and ecosystem health". *Nature* 484.7393, pp. 186–194.

Fisher, M. C., N. A. R. Gow, and S. J. Gurr (2016). „Tackling emerging fungal threats to animal health, food security and ecosystem resilience". *Philosophical Transactions of the Royal Society B: Biological Sciences* 371.1709, p. 20160332.

Fitzpatrick, D. a., M. E. Logue, J. E. Stajich, and G. Butler (2006). „A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis." *BMC Evolutionary Biology* 6, p. 99.

Flor, H. (1956). „The complementary genic systems in flax and flax rust". *Advances in Genetics* 8, pp. 29–54.

Floudas, D., M. Binder, R. Riley, K. Barry, R. A. Blanchette, B. Henrissat, A. T. Martinez, R. Otillar, J. W. Spatafora, J. S. Yadav, et al. (2012). „The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes." *Science* 336.6089, pp. 1715–1719.

Frick, W. F., J. F. Pollock, A. C. Hicks, K. E. Langwig, D. S. Reynolds, G. G. Turner, C. M. Butchkoski, and T. H. Kunz (2010). „An emerging disease causes regional population collapse of a common North American bat species." *Science* 329.5992, pp. 679–682.

Gamble, T., J. Coryell, T. Ezaz, J. Lynch, D. P. Scantlebury, and D. Zarkower (2015). „Restriction Site-Associated DNA Sequencing (RAD-seq) Reveals an Extraordinary Number of Transitions among Gecko Sex-Determining Systems". *Molecular Biology and Evolution* 32.5, pp. 1296–1309.

Gharib, W. H. and M. Robinson-Rechavi (2013). „The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC." *Molecular Biology and Evolution* 30.7, pp. 1675–1686.

Gladieux, P., A. Feurtey, M. E. Hood, A. Snirc, J. Clavel, C. Dutech, M. Roy, and T. Giraud (2015). „The population biology of fungal invasions". *Molecular Ecology* 24.9, pp. 1969–1986.

Gladieux, P., E. Vercken, M. C. Fontaine, M. E. Hood, O. Jonot, A. Couloux, and T. Giraud (2011). „Maintenance of Fungal Pathogen Species That Are Specialized to Different Hosts: Allopatric Divergence and Introgression through Secondary Contact." *Molecular Biology and Evolution* 28.1, pp. 459–471.

Goffeau, A, B. G. Barrell, H Bussey, R. W. Davis, B Dujon, H Feldmann, F Galibert, J. D. Hoheisel, C Jacq, M Johnston, et al. (1996). „Life with 6000 Genes". *Science* 274, pp. 546–567.

Gouveia, M. M. C., A. Ribeiro, V. M. P. Várzea, and C. J. Rodrigues (2005). „Genetic diversity in Hemileia vastatrix based on RAPD markers.“ *Mycologia* 97.2, pp. 396–404.

Goyeau, H., F. Halkett, M. F. Zapater, J. Carlier, and C. Lannou (2007). „Clonality and host selection in the wheat pathogenic fungus Puccinia triticina“. *Fungal Genetics and Biology* 44.6, pp. 474–483.

Grünwald, N., B. A. McDonald, and M. G. Milgroom (2016). „Population Genomics of Fungal and Oomycete Pathogens“. *Annual Review of Phytopathology* 54.1, pp. 323–346.

Hibbett, D. S., J. E. Stajich, and J. W. Spatafora (2013). „Toward genome-enabled mycology“. *Mycologia* 105.6, pp. 1339–1349.

Hickerson, M. J., B. C. Carstens, J Cavender-Bares, K. a. Crandall, C. H. Graham, J. B. Johnson, L Rissler, P. F. Victoriano, and a. D. Yoder (2010). „Phylogeography's past, present, and future: 10 years after Avise, 2000.“ *Molecular Phylogenetics and Evolution* 54.1, pp. 291–301.

Hindorf, H. and C. O. Omondi (2010). „A review of three major fungal diseases of Coffea arabica L. in the rainforests of Ethiopia and progress in breeding for resistance in Kenya“. *Journal of Advanced Research* 2.2, pp. 109–120.

Hohenlohe, P. A., P. C. Phillips, and W. A. Cresko (2011). „Using population genomics to detect selection in natural populations: Key concepts and methodological considerations“. *International Journal of Plant Science* 171.9, pp. 1059–1071.

Horner, D. S., G. Pavesi, T. Castrignanò, P. D. De Meo, S. Liuni, M. Sammeth, E. Picardi, and G. Pesole (2010). „Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing.“ *Briefings in bioinformatics* 11.2, pp. 181–197.

Hoskins, R., A. Phan, M Naeemuddin, F. Mapa, D. Ruddy, J. Ryan, L. Young, T Wells, C Kopczynski, and M. Ellis (2001). „Single Nucleotide Polymorphism Markers for Genetic Mapping in Drosophila melanogaster“. *Genome Research* 11, pp. 1100–1113.

Hou, Y., M. D. Nowak, V. Mirré, C. S. Bjorå, C. Brochmann, and M. Popp (2016). „RAD-seq data point to a northern origin of the arctic–alpine genus Cassiope (Ericaceae)“. *Molecular Phylogenetics and Evolution* 95, pp. 152–160.

Hudson, M. E. (2008). „Sequencing breakthroughs for genomic ecology and evolutionary biology.“ *Molecular Ecology Resources* 8.1, pp. 3–17.

Jarvis, E. D., S. Mirarab, A. J. Aberer, B. Li, P. Houde, C. Li, S. Y. W. Ho, B. C. Faircloth, B. Nabholz, J. T. Howard, et al. (2014). „Whole-genome analyses resolve early branches in the tree of life of modern birds.“ *Science* 346.6215, pp. 1320–1331.

Jeffroy, O., H. Brinkmann, F. Delsuc, and H. Philippe (2006). „Phylogenomics: the beginning of incongruence?“ *Trends in G5enetics* 22.4, pp. 225–231.

Jones, J. D. G. and J. L. Dangl (2006). „The plant immune system“. *Nature* 444.7117, pp. 323–329.

Kamvar, Z. N., J. F. Tabima, and N. J. Grünwald (2014). „Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction.“ *PeerJ* 2, e281.

Keller, I, C. E. Wagner, L Greuter, S Mwaiko, O. M. Selz, A Sivasundar, S Wittwer, and O Seehausen (2013). „Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes.“ *Molecular Ecology* 22.11, pp. 2848–2863.

Kiss, L., A. Pintye, G. M. Kovács, T. Jankovics, M. C. Fontaine, N. Harvey, X. Xu, P. C. Nicot, M. Bardin, J. a. Shykoff, et al. (2011). „Temporal isolation explains host-related genetic differentiation in a group of widespread mycoparasitic fungi." *Molecular Ecology* 20.7, pp. 1492–1507.

Kumar, S., A. J. Filipski, F. U. Battistuzzi, S. L. Kosakovsky Pond, and K. Tamura (2012). „Statistics and Truth in Phylogenomics". *Molecular Biology and Evolution* 29.2, pp. 457–472.

Kuramae, E. E., V. Robert, B. Snel, M. Weiss, and T. Boekhout (2006). „Phylogenomics reveal a robust fungal tree of life." *FEMS yeast research* 6, pp. 1213–1220.

Leaché, A. D. and J. R. Oaks (2017). „The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics". *Annual Review of Ecology, Evolution, and Systematics* 48.1, pp. 69–84.

Lindblad-Toh, K., E. Winchester, M. J. Daly, D. G. Wang, J. N. Hirschhorn, J. P. Laviolette, K. Ardlie, D. E. Reich, E. Robinson, P. Sklar, et al. (2000). „Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse". *Nature Genetics* 24.4, pp. 381–386.

Liu, Y., J. W. Leigh, H. Brinkmann, M. T. Cushion, N. Rodriguez-Ezpeleta, H. Philippe, and B. F. Lang (2009). „Phylogenomic analyses support the monophyly of taphrinomycotina, including Schizosaccharomyces fission yeasts". *Molecular Biology and Evolution* 26.1, pp. 27–34.

Longo, D. L. and J. M. Drazen (2016). „Data Sharing". *New England Journal of Medicine* 374.3, pp. 276–277.

Longo, G. and G. Bernardi (2015). „The evolutionary history of the embiotocid surfperch radiation based on genome-wide RAD sequence data". *Molecular Phylogenetics and Evolution* 88, pp. 55–63.

Louis, E. J. (2011). „Population genomics and speciation in yeasts". *Fungal Biology Reviews* 25.3, pp. 136–142.

Lowry, D. B., S. Hoban, J. L. Kelley, K. E. Lotterhos, L. K. Reed, M. F. Antolin, and A. Storfer (2017). „Breaking RAD: An evaluation of the utility of restriction site associated DNA sequencing for genome scans of adaptation". *Molecular Ecology Resources* 17.2, pp. 142–152.

Lu, A. and S. Guindon (2013). „Performance of Standard and Stochastic Branch-Site Models for Detecting Positive Selection among Coding Sequences." *Molecular Biology and Evolution* 31.2, pp. 484–495.

Ma, L.-J., J. W. Bennett, and N. D. Fedorova (2011). „Fungal Biology in the Age of Genomics". *Mycology* 2.3, p. 117.

Ma, T., J. Wang, G. Zhou, Z. Yue, Q. Hu, Y. Chen, B. Liu, Q. Qiu, Z. Wang, J. Zhang, et al. (2013). „Genomic insights into salt adaptation in a desert poplar." *Nature Communications* 4, p. 2797.

Maia, T. a., E. Maciel-Zambolim, E. T. Caixeta, E. S. G. Mizubuti, and L. Zambolim (2013). „The population structure of Hemileia vastatrix in Brazil inferred from AFLP". *Australasian Plant Pathology* 42.5, pp. 533–542.

Mardis, E. R. (2008a). „Next-generation DNA sequencing methods." *Annual Review of Genomics and Human Genetics* 9, pp. 387–402.

– (2008b). „The impact of next-generation sequencing technology on genetics". *Trends in Genetics* 24.3, pp. 133–141.

Markowetz, F. (2017). „All biology is computational biology". *PLoS Biology* 15.3, pp. 4–7.

Martin, F, D Cullen, D Hibbet, A Pisabarro, J. Spatafora, S. Baker, and I. Grigoriev (2011). „Sequencing the fungal tree of life". *New Phytologist* 190, pp. 818–821.

Mastretta-Yanes, A, N Arrigo, N Alvarez, T. H. Jorgensen, D Piñero, and B. C. Emerson (2015). „Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference“. *Molecular Ecology Resources* 15.1, pp. 28–41.

McCook, S. and J. Vandermeer (2015). „The Big Rust and the Red Queen: Long-Term Perspectives on Coffee Rust Research.“ *Phytopathology* 105.9, pp. 1164–1173.

McCormack, J. E., S. M. Hird, A. J. Zellmer, B. C. Carstens, and R. T. Brumfield (2013). „Applications of next-generation sequencing to phylogeography and phylogenetics“. *Molecular Phylogenetics and Evolution* 66.2, pp. 526–538.

Metzker, M. L. (2010). „Sequencing technologies - the next generation.“ *Nature Reviews Genetics* 11.1, pp. 31–46.

Milgroom, M. G. (1996). „Recombination and the multilocus structure of fungal populations.“ *Annual Review of Phytopathology* 34.1, pp. 457–477.

Misof, B., S. Liu, K. Meusemann, R. S. Peters, a. Donath, C. Mayer, P. B. Frandsen, J. Ware, T. Flouri, R. G. Beutel, et al. (2014). „Phylogenomics resolves the timing and pattern of insect evolution“. *Science* 346.6210, pp. 763–767.

Miyata, T. and T. Yasunaga (1980). „Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application“. *Journal of Molecular Evolution* 16.1, pp. 23–36.

Möller, M. and E. H. Stukenbrock (2017). „Evolution and genome architecture in fungal plant pathogens“. *Nature Reviews Microbiology* 15.12, pp. 756–771.

Neafsey, D. E., B. M. Barker, T. J. Sharpton, J. E. Stajich, D. J. Park, E. Whiston, C. Y. Hung, C. McMahan, J. White, S. Sykes, et al. (2010). „Population genomic sequencing of Coccidioides fungi reveals recent hybridization and transposon control“. *Genome Research* 20.7, pp. 938–946.

Nei, M. and T. Gojoborit (1986). „Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.“ *Molecular Biology and Evolution* 3.5, pp. 418–426.

Nekrutenko, A. and J. Taylor (2012). „Next-generation sequencing data interpretation: enhancing reproducibility and accessibility“. *Nature Reviews Genetics* 13.9, pp. 667–672.

Nemri, A., D. G. O. Saunders, C. Anderson, N. M. Upadhyaya, J. Win, G. J. Lawrence, D. a. Jones, S. Kamoun, J. G. Ellis, and P. N. Dodds (2014). „The genome sequence and effector complement of the flax rust pathogen Melampsora lini.“ *Frontiers in Plant Science* 5, p. 98.

Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song (2011). „Genotype and SNP calling from next-generation sequencing data“. *Nature Reviews Genetics* 12, pp. 443–451.

Nieuwenhuis, B. P. S. and T. Y. James (2016). „The frequency of sex in fungi“. *Philosophical Transactions of the Royal Society B: Biological Sciences* 371, p. 20150540.

Nishihara, H., N. Okada, and M. Hasegawa (2007). „Rooting the eutherian tree: the power and pitfalls of phylogenomics.“ *Genome Beiology* 8.9, R199.

Noronha-Wagner, M and A. J. Bettencourt (1967). „Genetic study of the resistance of coffea spp. to leaf rust“. *Canadian Journal of Botany* 45, pp. 2021–2031.

Nunes, C. C., L. A. Maffia, E. S. G. Mizubuti, S. H. Brommonschenkel, and J. C. Silva (2009). „Genetic diversity of populations of Hemileia vastatrix from organic and conventional coffee plantations in Brazil“. *Australian Plant Pathology* 38, pp. 445–452.

Orr, H. A. (2005). „The genetic theory of adaptation: a brief history“. *Nature Reviews Genetics* 6.2, pp. 119–127.

Pareek, C. S., R. Smoczynski, and A. Tretyn (2011). „Sequencing technologies and genome sequencing“. *Journal of Applied Genetics* 52.4, pp. 413–435.

Pavey, S. A., L. Bernatchez, N. Aubin-Horth, and C. R. Landry (2012). „What is needed for next-generation ecological and evolutionary genomics?“ *Trends in Ecology & Evolution* 27.12, pp. 673–678.

Pavey, S. A., J. Gaudin, E. Normandeau, M. Dionne, M. Castonguay, C. Audet, and L. Bernatchez (2015). „RAD Sequencing Highlights Polygenic Discrimination of Habitat Ecotypes in the Panmictic American Eel“. *Current Biology* 25.12, pp. 1666–1671.

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra (2012). „Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species“. *PLoS ONE* 7.5, e37135.

Philippe, H. and M. Blanchette (2007). „Overview of the First Phylogenomics Conference“. *BMC Evolutionary Biology* 7, S1.

Philippe, H. H., F. F. Delsuc, H. Brinkmann, and N. Lartillot (2005). „Phylogenomics“. *Annual Review of Ecology, Evolution, and Systematics* 36.1, pp. 541–562.

Pop, M. and S. L. Salzberg (2008). „Bioinformatics challenges of new sequencing technology.“ *Trends in Genetics* 24.3, pp. 142–149.

Posada, D. (2016). „Phylogenomics for Systematic Biology“. *Systematic Biology* 65.3, pp. 353–356.

Prakash, N., J Devasia, K Das Divya, B. Manjunatha, H. Seetharam, A Kumar, and Jayarama (2014). „Breeding for rust resistance in Arabica – where we are and what next?“ In: *Proceedings of the 25th International Conference on Coffee Science (ASIC)*, B10.

Pyron, R. A. (2015). „Post-molecular systematics and the future of phylogenetics“. *Trends in Ecology & Evolution* 30.7, pp. 384–389.

Ritschel, A (2005). „Monograph of the genus Hemileia (Uredinales)“. In: *Bibliotheca Mycologica*. Ed. by A Bresinsky, H Butin, and P Tudzinski. Stuttgart: J. Cramer, pp. 3–132.

Robbertse, B., J. B. Reeves, C. L. Schoch, and J. W. Spatafora (2006). „A phylogenomic analysis of the Ascomycota.“ *Fungal genetics and biology* 43, pp. 715–725.

Rochette, N. C., C. Brochier-Armanet, and M. Gouy (2014). „Phylogenomic Test of the Hypotheses for the Evolutionary Origin of Eukaryotes.“ *Molecular Biology and Evolution* 31.4, pp. 832–845.

Rozo, Y., C. Escobar, Á. Gaitán, and M. Cristancho (2012). „Aggressiveness and Genetic Diversity of Hemileia vastatrix During an Epidemic in Colombia“. *Journal of Phytopathology* 160.11-12, pp. 732–740.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, et al. (2001). „A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms“. *Nature* 409.6822, pp. 928–933.

Schie, C. C. van and F. L. Takken (2014). „Susceptibility Genes 101: How to Be a Good Host“. *Annual Review of Phytopathology* 52.1, pp. 551–581.

Schuster, S. C. (2008). „Next-generation sequencing transforms today's biology“. *Nature Methods* 5.1, pp. 16–18.

Sedghifar, A., P. Saelao, and D. J. Begun (2016). „Genomic patterns of geographic differentiation in Drosophila simulans“. *Genetics* 202.March, pp. 1229–1240.

Shaffer, H. B., P. Minx, D. E. Warren, A. M. Shedlock, R. C. Thomson, N. Valenzuela, J. Abramyan, C. T. Amemiya, D. Badenhorst, K. K. Biggar, et al. (2013). „The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage.“ *Genome Biology* 14.3, R28.

Silva, M., V Várzea, L Guerra-Guimarães, H Azinheira, D Fernandez, A.-S. Petitot, B Bertrand, P Lashermes, and M Nicole (2006). „Coffee resistance to the main diseases: leaf rust and coffee berry disease“. *Brazilian Journal of Plant Physiology* 18, pp. 119–147.

Slot, J. C. and A. Rokas (2011). „Horizontal transfer of a large and highly toxic secondary metabolic gene cluster between fungi“. *Current Biology* 21.2, pp. 134–139.

Snel, B., M. a. Huynen, and B. E. Dutilh (2005). „Genome trees and the nature of genome evolution.“ *Annual Review of Microbiology* 59, pp. 191–209.

Sovic, M. G., B. C. Carstens, and H. L. Gibbs (2016). „Genetic diversity in migratory bats: Results from RADseq data for three tree bat species at an Ohio windfarm“. *PeerJ* 4, e1647.

Spanu, P. D. (2012). „The Genomics of Obligate (and Nonobligate) Biotrophs“. *Annual Review of Phytopathology* 50.1, pp. 91–109.

Spanu, P. D., J. C. Abbott, J. Amselem, T. A. Burgis, D. M. Soanes, K. Stüber, E. Ver Loren van Themaat, J. K. M. Brown, S. A. Butcher, S. J. Gurr, et al. (2010). „Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism.“ *Science* 330.6010, pp. 1543–1546.

Staples, R. C. (2000). „Research on the Rust Fungi During the Twentieth Century“. *Annual Review of Phytopathology* 38.1, pp. 49–69.

Stinchcombe, J. R. and H. E. Hoekstra (2007). „Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits“. *Heredity* 100.2, pp. 158–170.

Stukenbrock, E. H. and T. Bataillon (2012). „A Population Genomics Perspective on the Emergence and Adaptation of New Plant Pathogens in Agro-Ecosystems“. *PLoS Pathogens* 8.9, e1002893.

Stukenbrock, E. H. and B. A. McDonald (2008). „The origins of plant pathogens in agro-ecosystems.“ *Annual Review of Phytopathology* 46, pp. 75–100.

Talhinhas, P., H. G. Azinheira, B. Vieira, A. Loureiro, S. Tavares, D. Batista, E. Morin, A.-S. Petitot, O. S. Paulo, J. Poulain, et al. (2014). „Overview of the functional virulent genome of the coffee leaf rust pathogen Hemileia vastatrix with an emphasis on early stages of infection“. *Frontiers in Plant Science* 5, p. 88.

Talhinhas, P., D. Batista, I Diniz, A Vieira, D. Silva, A Loureiro, S Tavares, A. Pereira, H. Azinheira, L Guerra-Guimarães, et al. (2017). „Pathogen profile The coffee leaf rust pathogen Hemileia vastatrix : one and a half centuries around the tropics“. *Molecular Plant Pathology* 18.8, pp. 1039–1051.

Tavares, S., A. P. Ramos, A. S. Pires, H. G. Azinheira, P. Caldeirinha, T. Link, R. Abranches, M. D. C. Silva, R. T. Voegele, J. Loureiro, et al. (2014). „Genome size analyses of Pucciniales reveal the largest fungal genomes.“ *Frontiers in Plant Science* 5, p. 422.

The Arabidopsis Genome Initiative (2000). „Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.“ *Nature* 408.6814, pp. 796–815.

Thomson, R. C., I. J. Wang, and J. R. Johnson (2010). „Genome-enabled development of DNA markers for ecology, evolution and conservation." *Molecular Ecology* 19.11, pp. 2184–2195.

Toonen, R. J., J. B. Puritz, Z. H. Forsman, J. L. Whitney, I. Fernandez-Silva, K. R. Andrews, and C. E. Bird (2013). „ezRAD: a simplified method for genomic genotyping in non-model organisms". *PeerJ* 1, e203.

Torruella, G., A. de Mendoza, X. Grau-Bové, M. Antó, M. Chaplin, J. del Campo, L. Eme, G. Pérez-Cordón, C. Whipps, K. Nichols, et al. (2015). „Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi". *Current Biology*, pp. 2404–2410.

Vamathevan, J. J., S. Hasan, R. D. Emes, H. Amrine-Madsen, D. Rajagopalan, S. D. Topp, V. Kumar, M. Word, M. D. Simmons, S. M. Foord, et al. (2008). „The role of positive selection in determining the molecular cause of species differences in disease." *BMC Evolutionary Biology* 8, p. 273.

Van der Merwe, M., M. Kinnear, L. Barrett, P. Dodds, L Ericson, P. Thrall, and J. Burdon (2009). „Positive selection in AvrP4 avirulence gene homologues across the genus Melampsora." *Proceedings of the Royal Society B Biological Sciences* 276.1669, pp. 2913–2922.

Venter, J. C., M. D. D. Adams, E. W. W. Myers, P. W. W. Li, R. J. J. Mural, G. G. G. Sutton, H. O. O. Smith, M. Yandell, C. A. A. Evans, R. A. A. Holt, et al. (2001). „The sequence of the human genome". *Science* 291.5507, pp. 1304–1351.

Via, A., T. Blicher, E. Bongcam-Rudloff, M. D. Brazas, C. Brooksbank, A. Budd, J. De Las Rivas, J. Dreyer, P. L. Fernandes, C. Van Gelder, et al. (2013). „Best practices in bioinformatics training for life scientists". *Briefings in Bioinformatics* 14.5, pp. 528–537.

Vitti, J. J., S. R. Grossman, and P. C. Sabeti (2013). „Detecting Natural Selection in Genomic Data". *Annual Review of Genetics* 47.1, pp. 97–120.

Voegele, R. T. and K. W. Mendgen (2011). „Nutrient uptake in rust fungi: How sweet is parasitic life?" *Euphytica* 179.1, pp. 41–55.

Vogel, K. J. and N. a. Moran (2013). „Functional and evolutionary analysis of the genome of an obligate fungal symbiont." *Genome Biology and Evolution* 5.5, pp. 891–904.

Wang, S., E. Meyer, J. K. Mckay, and M. V. Matz (2012). „2b-RAD: A simple and flexible method for genome-wide genotyping". *Nature Methods* 9.8, pp. 808–810.

Waterston, R. H., K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, et al. (2002). „Initial sequencing and comparative analysis of the mouse genome". *Nature* 420.6915, pp. 520–562.

Webb, J. (2002). *Tropical pioneers human agency and ecological change in the highlands of Sri Lanka, 1800–1900*. Athens: Ohio University Press.

Wellenreuther, M. and B. Hansson (2016). „Detecting Polygenic Evolution: Problems, Pitfalls, and Promises". *Trends in Genetics* 32.3, pp. 155–164.

Yang, Z and J. Bielawski (2000). „Statistical methods for detecting molecular adaptation." *Trends in Ecology & Evolution* 15.12, pp. 496–503.

Yang, Z. (2005). „The power of phylogenetic comparison in revealing protein function." *Proceedings of the National Academy of Sciences of the United States of America* 102.9, pp. 3179–3180.

– (2007). „PAML 4: phylogenetic analysis by maximum likelihood." *Molecular Biology and Evolution* 24.8, pp. 1586–1591.

– (2011). *A Primer of Probability & Statistics*. Tech. rep. September, p. 43.

Yang, Z. and R. Nielsen (2002). „Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages." *Molecular Biology and Evolution* 19.6, pp. 908–917.

– (2008). „Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage." *Molecular Biology and Evolution* 25.3, pp. 568–579.

Zambolim, L. (2016). „Current status and management of coffee leaf rust in Brazil". *Tropical Plant Pathology* 41.1, pp. 1–8.

Zapata, F., N. G. Wilson, M. Howison, S. C. S. Andrade, K. M. Jorger, M. Schrodl, F. E. Goetz, G. Giribet, and C. W. Dunn (2014). „Phylogenomic analyses of deep gastropod relationships reject Orthogastropoda". *Proceedings of the Royal Society B: Biological Sciences* 281.1794, pp. 20141739–20141739.

Zhan, J., P. H. Thrall, and J. J. Burdon (2014). „Achieving sustainable plant disease management through evolutionary principles". *Trends in Plant Science* 19.9, pp. 570–575.

Zhan, J., P. H. Thrall, J. Papaïx, L. Xie, and J. J. Burdon (2015). „Playing on a Pathogen's Weakness: Using Evolution to Guide Sustainable Plant Disease Control Strategies". *Annual Review of Phytopathology* 53.1, pp. 19–43.

Zhang, J., R. Chiodini, A. Badr, and G. Zhang (2011). „The impact of next-generation sequencing on genomics." *Journal of Genetics and Genomics* 38.3, pp. 95–109.

2

Genomic patterns of positive selection at the origin of rust fungi

Diogo N. Silva [1,2,3], Sebastien Duplessis [4,5], Pedro Talhinhas [1,6], Helena Azinheira [1,6], Octávio S. Paulo [2], Dora Batista [1,2]

[1] Centro de Investigação das Ferrugens do Cafeeiro, Instituto Superior de Agronomia, Universidade de Lisboa, Oeiras, Portugal.

[2] Computational Biology and Population Genomics group, cE3c – Centre for Centre for Ecology Evolution and Environmental Changes, Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal.

[3] Departamento de Biologia e CESAM - Centro de Estudos do Ambiente e do Mar, Universidade de Aveiro, Aveiro, Portugal.

[4] Institut National de la Recherche Agronomique, UMR1136 INRA/Université de Lorraine Interactions Arbres - Microorganismes, Champenoux, France.

[5] Université de Lorraine,UMR1136, INRA/Université de Lorraine Interactions Arbres - Microorganismes, Vandoeuvre-lés- Nancy, France.

[6] LEAF, Linking Landscape, Environment, Agriculture and Food, Instituto Superior de Agronomia, Universidade de Lisboa, Lisbon, Portugal.

abovecaptionskip

2.1 Abstract

Understanding the origin and evolution of pathogenicity and biotrophic life-style of rust fungi has remained a conundrum for decades. Research on the molecu-

Published in PLoS One. DOI: 10.1371/journal.pone.0143959

lar mechanisms responsible for rust fungi evolution has been hampered by their biotrophic life-style until the sequencing of some rust fungi genomes. With the availability of multiple whole genomes and EST data for this group, it is now possible to employ genome-wide surveys and investigate how natural selection shaped their evolution. In this work, we employed a phylogenomics approach to search for positive selection and genes undergoing accelerated evolution at the origin of rust fungi on an assembly of single copy genes conserved across a broad range of basidiomycetes. Up to 985 genes were screened for positive selection on the phylogenetic branch leading to rusts, revealing a pervasive signal of positive selection throughout the data set with the proportion of positively selected genes ranging between 19.6-33.3%. Additionally, 30 genes were found to be under accelerated evolution at the origin of rust fungi, probably due to a mixture of positive selection and relaxation of purifying selection. Functional annotation of the positively selected genes revealed an enrichment in genes involved in the biosynthesis of secondary metabolites and several metabolism and transporter classes.

2.2 Introduction

During the past 30 years, the Neutralist-Selectionist debate has gradually settled into a consensus that both neutral drift and selection are fundamental pillars in evolution, with positive darwinian selection playing a major role in evolutionary change (Anisimova and Liberles, 2007; Vitti et al., 2013). The action of positive selection on coding regions traditionally manifests itself by an excess of non-synonymous substitutions relatively to synonymous substitutions (Yang and Bielawski, 2000). The search for such signatures within the genomes of organisms has been an exciting and fruitful pursuit of modern biologists, since adaptive changes in genes are ultimately responsible for evolutionary innovations that can have a significant impact on the survival and adaptation of species to their environment and generate a breadth of phenotypic diversity (MacColl, 2011). Therefore, there is a growing interest in answering long-standing questions such as which genes were affected by positive selection, how strong was the effect, and when did it occur. Pursuing this endeavor is the first crucial step towards understanding the consequences of genetic

variation on phenotypes and ultimately, its contribution to the fitness of individuals (Barrett and Hoekstra, 2011; MacColl, 2011; Vitti et al., 2013).

Plant pathogens have been obvious targets of studies aimed at finding molecular fingerprints of positive selection on avirulence genes, elicitors and effectors involved in arms-race interactions between hosts and pathogens (Tellier et al., 2014). Through this approach, several genes previously suspected to be involved in pathogenicity were also found to be under positive selection, such as the SnToxA toxin in *Pyrenophora* spp. (Stukenbrock and McDonald, 2007), the RXLR effectors of oomycetes (Win et al., 2007), or the *AvrP4* avirulence gene of *Melampsora* spp. (Merwe et al., 2007). These studies were important to confirm and better understand how these genes affected pathogenicity, but their range was quite limited by having to rely only on previously identified candidate genes. More recently, with the increasing availability of whole-genome sequences and sophistication of statistical methods, researchers became able to harness information from thousands of genes across multiple species and to perform genome-scans for positive selection that forego the need for *a priori* knowledge on specific genes (Roux et al., 2014). With these advances, finding and understanding natural selection transited from hypothesis-testing to hypothesis-generating science (Vitti et al., 2013). By comparing several ortholog sequences across multiple species and employing a maximum likelihood test of d_N/d_S ratios, such "blind" approaches have the enormous potential of pinpointing not only genes or even amino acid residues targeted by positive selection that were not previously considered, but also the functional categories enriched for positively selected genes following a relevant evolutionary event or transition (Aguileta et al., 2010). Moreover, it is now possible to uncover episodic positive selection acting on a pre-specified branch of the phylogenetic tree and, therefore, on a specific evolutionary time (Yang and Nielsen, 2002; Zhang et al., 2005). The power and usefulness of these genome scans led to their application in a wide range of taxa to investigate a number of questions, such as differential adaptation to disease in humans and chimpanzees (Vamathevan et al., 2008), salt adaptation in a desert poplar (Ma et al., 2013), extreme physiological adaptations in turtles (Shaffer et al., 2013), among other examples (Nery et al., 2013; Wang et al., 2013;

Zhao et al., 2013). However, their application in pathogenic fungal systems has been more limited (Aguileta et al., 2010; Aguileta et al., 2012).

In this work, a genome-wide scan for positive selection and genes undergoing accelerated evolution integrated into a phylogenomics framework was used to investigate the evolutionary origin of rust fungi (Basidiomycota, Pucciniales), namely regarding the distinctive features of their biotrophic life-style and pathogenicity. Rust fungi are a diverse and economically important group of obligate plant pathogens that cause devastating diseases on cultivated plants of almost all taxonomic families (Staples, 2000; Fernandez et al., 2013). Despite their importance, the obligate biotrophic life-style and inability to grow axenically has complicated the research on the molecular mechanisms responsible for the evolution of their pathogenicity. Therefore, the molecular features underlying the adaptations of obligate biotrophic associations with host plants remained largely unknown until the publication of the first two rust genome sequences, *Melampsora larici-populina* and *Puccinia graminis* f. sp. *tritici* (Duplessis et al., 2011). These resources provided key insights to understand the origin and singularity of rust fungi such as: (i) expansion of lineage-specific gene families that account for a high number of predicted genes compared to other basidiomycete pathogens; (ii) absence of sucrose transporters as well as loss of some genes involved in inorganic nitrogen and sulphur uptake and assimilation; (iii) a reduced carbohydrate active enzymes repertoire, albeit some classes may be expanded; (iv) larger repertoire of small secreted proteins, most of which are also up regulated transcripts and lineage specific; and (v) transposon proliferation.

Arguably, the discovery of these genome-scale changes significantly advanced our knowledge on rust fungi, focusing on new genomic features that are unique to rust genomes (Duplessis et al., 2011; Cantu et al., 2013; Nemri et al., 2014; Pendleton et al., 2014). Considerably less explored remains the role of adaptive genetic variation on genetic material shared between rust fungi and other Basidiomycetes, particularly on the phylogenetic root branch of the Puccinales. When tackling the issue of the adaptations that took place on the origin of rust fungi and their life-style, the importance of investigating variation on genes conserved across multiple species is twofold: first, it has been extensively demonstrated that even large changes of

form and function can evolve by altering the sequence and functionality of conserved proteins (Anisimova and Liberles, 2007; Carroll, 2008); second, the genetic material of the common ancestor of rust fungi had to be shared to some extent with several contemporary populations and species, as it occurs for the common ancestors of all organisms (Benner et al., 2007). Therefore, the first adaptive changes would have been required to occur in the shared genetic material (Barrett and Schluter, 2008). In this work, by using sophisticated and powerful methods capable of detecting episodic positive selection and by including a more basal lineage of the Pucciniales, it was our goal to identify genes with signatures of positive selection specifically at the time when the rust fungi originated and provide insights on the strength and pervasiveness of positive selection in rust genomes.

Our specific aims were thus to: (i) detect and identify the largest number of single-copy orthologs shared among a data set of 67 Basidiomycota and Ascomycota species with a combination of genomic and Expressed Sequence Tags (EST) data; (ii) screen for episodic positive selection acting on specific amino acids and determine the magnitude of the signal for positive selection acting on the root of the rust fungi ; (iii) detect genes that are significantly accelerated on the root of the rust fungi; and (iv) annotate the candidate genes and investigate if certain functional classes are enriched for positively selected genes.

2.3 Materials and Methods

2.3.1 Genomic and EST data

Genomic data was collected from 37 Basidiomycota species from public databases, as well as from 9 Ascomycota species to use as outgroups. This data set included the complete genome of three rust fungi (*Melampsora larici-populina*, *Puccinia graminis* f. sp. *tritici* and *P. triticina*). Additionally, EST data from 65 Basidiomycota species was gathered from the NCBI public repository, except for *Hemileia vastatrix* (Talhinhas et al., 2014). In total, 753 848 EST sequences were compiled. The full

list of species and their corresponding genomic sources and citations is provided in Table A.1.1.

2.3.2 Processing of EST data

Before processing the EST data, species' sequences for which complete genome sequences were available in alternative were removed. Twenty one species were thus excluded, resulting in an initial data set with 44 species and 259 704 sequences. ESTs were screened for vector contaminants and trimmed using SeqClean (http://seqclean.sourceforge.net/) (Chen et al., 2007) against the UniVec database (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html), and sequences with less than 100 bp were discarded. Interspersed repeats and low complexity regions in the ESTs were then masked with RepeatMasker (A.F.A. Smit, R. Hubley & P. Green RepeatMasker at http://repeatmasker.org). To this end, the RMBlast search engine and the repeat database RepBase, which is optimized for RepeatMasker, were used. These cleaning steps resulted in a final data set of 255 156 sequences, with 51 192 Kbp trimmed and 2 080 Kbp masked.

2.3.3 Ortholog search strategy

The detection of ortholog sequences in the combined genomic and EST data sets followed two sequential approaches. First, orthology relationships among 488 087 protein sequences for the 37 Basidiomycota species with complete draft genomes were assessed using the OrthoMCL software (Li et al., 2003) via the orthomcl_pipeline.py script (hosted at https://github.com/ODiogoSilva/Silva2014_RustPhylogenomics). The procedure starts with all-vs-all BLASTp comparisons, for which an e-value cut-off of $1e^{-5}$ was chosen. Then, a Markov Clustering (MCL) algorithm (http://micans.org/mcl/) (Dongen, 2000; Enright et al., 2002) was used to create clusters of putative orthologs and co-orthologs according to an inflation parameter that controls cluster tightness. To strike a balance between cluster tightness and the probability of breaking up clusters with the same orthologs, the intermediate inflation value of 3 was used. Since the output of OrthoMCL con-

tains recent paralogs in addition to single copy genes, clusters with more than one gene copy per species or comprising less than 9 species out of the 37 total species were removed. Furthermore, we only included single-copy genes represented by at least one species of the Pucciniales. This resulted in a final data set of 1 715 core single-copy ortholog clusters, 1 250 of which contained the three Pucciniales (Tables A.1.2 to A.1.8 in pages 190 to 208). Nine additional Ascomycota species were added for outgroup purposes. For future reference in this paper, ortholog clusters will be simply referred to as genes.

Second, the taxonomic coverage of the previously determined core single-copy genes was then expanded with EST data using the software **HaMSTr** (Ebersberger et al., 2009). By complementing only this set of genes, we attempted to substantially reduce the chance of inadvertently introducing paralogs from the EST data. Each core gene was aligned in **MAFFT** v7.0 (Katoh and Standley, 2013) using the `L-INS-i` method and Hidden Markov Model profiles were then constructed for each alignment using `hmmbuild`, included in the **HMMR3** package (`http://hmmer.janelia.org`) (Eddy, 2010). All 37 Basidiomycota species were used as representatives during the **HaMSTr** searches and the "`-relaxed`" option of the program was specified, in order to relax the constraint of a potential ortholog to be present in all representative species. Finally, species whose data could not complement at least 50 core genes were discarded, resulting in a final inclusion of 21 species with EST sequences. Therefore, the largest data set assembled contained genomic data from 37 Basidiomycota species, 9 Ascomycota species and EST data from 21 additional species, for a total of 67 species.

2.3.4 Sequence alignment and filtering

Genes were aligned with **MAFFT** v7.0 using the `L-INS-i` method. Alignment columns with excessive missing data were filtered using **TriFusion** v0.1 software (`https://github.com/ODiogoSilva/TriFusion`), which removed columns with a proportion of missing data above 50% in the extremities of the alignment and columns with a proportion of gaps and missing data above 75%. In order to take the alignment uncertainty into account when performing the phylogenomic recon-

struction, weights were attributed to each alignment column, using the probabilistic framework implemented in **ZORRO** (Wu et al., 2012), for latter interpretation by the phylogenetic reconstruction software. For the detection of positive selection, the DNA sequences corresponding to each species in the protein alignments were assembled and aligned with **TranslatorX** (Abascal et al., 2010) using the protein alignment as reference. Since the software used to detect positive selection does not incorporate the probabilistic weighting schemes from **ZORRO**, filtering of fast evolving and potentially misaligned alignment blocks was performed following two independent approaches: **Guidance** (Sela et al., 2015), using default parameters and **MAFFT** as the multiple sequence alignment algorithm, and **Gblocks** (Talavera and Castresana, 2007), within the framework of **TranslatorX**, using default parameters, except that columns with half gap positions were allowed (option `-b5=h`).

2.3.5 Data set assembly

In total, six data sets were assembled for this study with specific aims (Table 2.4.1). The first four data sets were used for phylogenetic reconstructions and contained only protein sequences. Two of these were composed of sequence data from 37 Basidiomycetes plus 9 Ascomycetes with complete genomes and they differed only on the minimum amount of species required for each gene: the `genomic46sp_sparse` data set required a minimum of 9 Basidiomycota species to be represented, while the `genomic46sp_dense` data set required at least 36 Basidiomycota species to be represented. The other two protein data sets extended the taxonomic coverage of the previous data sets with EST data from 21 additional species: the `combined67sp_sparse` extends the genes from `genomic46sp_sparse`, and `combined67sp_dense` extends the genes from `genomic46sp_dense`. The two final data sets were used for positive selection detection analyses and contained only nucleotide sequences derived from the genes in the `genomic46sp_sparse` data set but without the Ascomycota species and with the additional filter that at least one *Puccinia* spp. and *M. larici-populina* must be present: the `basidioPAML` data set contained sequences obtained only from complete genomes while the `basidioPAML_Hv` data set contained a subset of the sequences in `basidioPAML` for which EST data of *H. vastatrix* was available. The

name of the alignments on each data set is composed of the *BasidioOnly* string, followed by an arbitrary numeric identified (*e.g.*: *BasidioOnly1001*).

2.3.6 Phylogenomic reconstruction

Maximum likelihood (ML) tree reconstruction was undertaken using **RAxML** v8.0.9 (Stamatakis, 2014) and the `PROTGAMMALG` model of sequence evolution for the protein data sets (Table 2.4.1). Node support was estimated by performing 250 non-parametric bootstrap replicates. ML searches and bootstrap replicates were performed in the CIPRES Science Gateway clusters (Miller et al., 2010).

2.3.7 Detection of positive selection at the origin of the Pucciniales

Prior to the detection of positive selection, sequence alignments showing a significant amount of substitution saturation were identified and removed from further analyses using an information entropy-based index implemented in **DAMBE5** (Xia et al., 2003; Xia, 2013) with a p-value threshold of 0.05. In alignments containing gaps, only gap free zones were analyzed. The branch-site test of positive selection implemented in **PAML** v4.4 (Zhang et al., 2005; Yang, 2007) was then used to test for positive selection in a specific branch of the phylogenetic tree by estimating the ratio of synonymous and nonsynonymous mutations, ω. This method detects signatures of positive selection when there is an excess of non-synonymous mutations relative to synonymous mutations on any given codon and a given phylogenetic branch. Two specific branches of interest, called foreground branches, were considered in this study: (i) the one on the origin of all included Pucciniales, containing the ancient lineage of *H. vastatrix* (`basidioPAML_Hv`), and (ii) the one on the origin of the majority of Pucciniales with draft genomes used in this study (`basidioPAML`) (Figure 2.1).

By studying these two foreground branches it will be possible to assess the impact of including or excluding a basal lineage of a taxonomic group when investigating its origin, and will allow an assessment of the impact that introducing lower quality EST

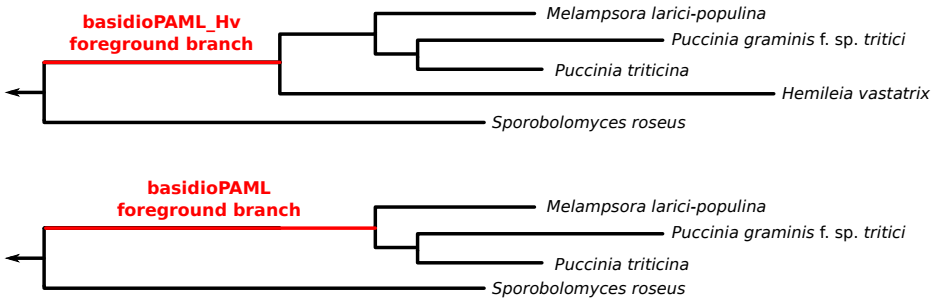

Figure 2.1. Schematic representation of the different foreground branches being considered in the `basidioPAML` and `basidioPAML_Hv` data sets.

data may have on the results. The consensus topology of the ML tree obtained from the RAxML analyses was used in both analyses. After performing the statistical tests, sites with identical amino acid residues but divergent codons that were found to be under positive selection were discarded in order to produce a final data set containing only positively selected sites that resulted in an amino acid change (a detailed description and discussion of the action of positive selection on conserved amino acid sites is presented in section 2.6.1).

Computations were performed with SLIMCODEML (Valle et al., 2014), a faster version of CODEML optimized for the branch-site model, using input alignments filtered either by GUIDANCE or GBLOCKS. The output of the alternative and null models for each alignment was processed with the custom pipeline CODEML_PARSER.PY (https://github.com/ODiogoSilva/Silva2014_RustPhylogenomics), which, among other tasks, retrieved the lnL scores, calculated the log-likelihood ratios and compared them to a χ^2 distribution with one degree of freedom. To correct for multiple testing, a False Discovery Rate (FDR) correction was performed and the alternative models were considered to be significantly better than the respective null model when $q < 0.05$.

2.3.8 Assessment of gene molecular rates

In parallel to the genome-wide screening of positive selection, the evolutionary rate of the genes with highest taxa representation in the `genomic46_dense` data set, was assessed for several branches of the Pucciniales by fitting a relaxed molecular clock

to the data using the MCMCTree software, included in the PAML v4.7 package (Yang, 2007). The phylogenetic tree resulting from the phylogenomic analyses was used as input for all alignments. Since not all species were present in all alignments, the input phylogenetic tree for each alignment was pruned using DendroPy (https://www.dendropy.org/) according to the present taxa, with the condition that all representative species of the Pucciniales had to be included. If this condition was not verified, the alignment was discarded from the analysis. Since the approximate method of MCMCTree was mandatory for protein alignments, maximum likelihood estimates of branch lengths, the gradient vector and Hessian matrix were initially calculated using the Codeml program using the WAG+Γ4 substitution model. Then, these calculations were used by MCMCTree to estimate divergence times and substitution rates on the tree topology. The calibration point on the root of the input phylogenetic tree was fixed to an arbitrary value of 5 (i.e., the prior was a uniform distribution with [5,5] bounds). Each analysis of an individual alignment was run twice to ensure convergence of the results and the mean of the gamma prior on the overall substitution rate was verified in each case to ensure that it was of the same magnitude as the estimated substitution rate. To assess whether a certain gene is evolving at a faster rate in the ancestral branch of all rust fungi, the Grubb's test for outliers was used. This statistical test is able to detect outlier values in a univariate data set. Therefore, genes were considered to be evolving at an atypical substitution rate at the root branch of the Pucciniales when the test statistic, Grubbs G, for that value was below a rejection cut-off of 0.05.

2.3.9 Functional annotation

To gain insight on the functions of both the putative positively selected and genes under accelerated evolution, a functional annotation was performed using the eu-Karyotic Orthologous Group (KOG) terminology, according to the eggNOG 4.0 (Powell et al., 2014) database and using its BLAST-based online search tool (http://eggnog.embl.de). Additionally, a Gene Ontology enrichment analysis was performed to determine if any category was overrepresented for genes under positive selection and faster evolutionary rate. Statistically significant enrichment was tested against a reference of all genes analysed using the Fisher's exact test and a

p-value for the independence of rows and columns in a 2×2 contingency table was computed. Significance was considered for $p < 0.10$.

To collect a more detailed information on those genes, sequence homology searches were performed against several databases: the NCBI non-redundant (`nr`) nucleotide and protein databases (`www.ncbi.nlm.nih.gov`), the genome sequences of *Melampsora larici-populina*, *Puccinia graminis* f. sp. *tritici* and *Uromyces fabae* (Duplessis et al., 2011; Link et al., 2014) and the Pathogen-Host Interaction (PHI-base v3.2) reference database (`www.phi-base.org`). **BLAST** searches were conducted with an e-value cut-off of $1e^{-5}$, and only the best hit was considered.

2.4 Results

2.4.1 Assembly of phylogenomic data sets

Six data sets were assembled for this study with different purposes and information concerning the number of genes, species and alignment columns is provided in (Table 2.4.1). The two data set pairs `genomic46sp_sparse` / `combined67sp_sparse` and `genomic46sp_dense` / `combined67sp_dense`, were used for phylogenomic analyses, the data set `genomic46sp_dense` was used for the evolutionary rate analysis and the data sets `basidioPAML` and `basidioPAML_Hv` were used for the positive selection detection analyses. Information about missing data and average gene length for each data set is provided from Tables A.1.2 to A.1.8 in pages 190 to 208.

Table 2.4.1. Description of the data sets assembled, including the number of genes, species alignment columns and the analysis performed.

Data set	Data type	Genes	Species	AC[1]	EST?	Analysis
`combined67sp_sparse`	Protein	1 715	68	775 565	yes	RaxML
`combined67sp_dense`	Protein	614	68	326 303	yes	RaxML
`genomic46sp_sparse`	Protein	1 715	47	842 795	no	RaxML
`genomic46sp_dense`	Protein	614	47	350 535	no	RaxML; MCMCTree
`basidioPAML`	DNA	985	9-36	45-3432	no	PAML
`basidioPAML_Hv`	DNA	531	14-37	15-3015	yes	PAML

[1] Alignment columns

For the phylogenomic reconstruction, the two pairs of data sets composed of protein sequences with and without translated EST data were used to assess evolutionary relationships and the potential impact of EST sequences on the phylogeny: (i) genomic46sp_sparse / combined67sp_sparse data sets contained 1 715 genes with at least 9 Basidiomycota species represented; and (ii) genomic46sp_dense / combined67sp_dense data sets comprised 614 genes with a minimum of 36 Basidiomycota species represented. The data sets that contained EST data presented an expected high level of missing amino acid data (combined67sp_sparse: mean = 42.0%, standard deviation = 36.2; combined67sp_dense: mean = 34.49%, standard deviation = 38.2) in great part due to the fragmented and limited distribution of EST sequences across the core genes. In contrast, the average missing amino acid data of the corresponding genome-only pairs was relatively low according to phylogenomic standards (genomic46sp_sparse: mean = 27.4%, standard deviation = 19.0; genomic46sp_dense: mean = 13%, standard deviation = 10.7). For the assessment of evolutionary rates, only genes from the genomic46sp_dense data set were used, since the higher sequence error rate of EST sequences could artificially increase estimates of substitution rate and produce misleading results. Concerning the two data sets composed of nucleotide sequences, the basidioPAML data set was composed of 985 genes from the genomic46sp_dense data set, with at least one *Puccinia* spp. and *M. larici-populina* represented and after the exclusion of 21 genes with significant substitution saturation. In an effort to include an ancient lineage of the Pucciniales, the previous data set was complemented with EST data of *H. vastatrix*. The basidioPAML_Hv data set contained 531 out of the 985 genes for which sequence data of *H. vastatrix* was available and after the exclusion of 11 saturated genes. The average proportion of missing genes per species was low in both data sets (basidioPAML: mean = 5.0%, standard deviation = 6.0; basidioPAML_Hv: mean = 3.9%, standard deviation = 4.8) as well as the proportion of missing nucleotides in the alignments (basidioPAML: mean = 11.8%, standard deviation = 5.6; basidioPAML_Hv: mean = 4.9%, standard deviation = 8.0). The average sequence length was also similar between data sets (basidioPAML: mean = 736, standard deviation = 12; basidioPAML_Hv: mean = 825, standard deviation = 64). Even though the average proportion of missing data was smaller in the basidioPAML_Hv data set, the sequence data of *H. vastatrix* presented a substantial amount of missing data

(45.8%), mainly due to the presence of gaps in the alignments, that also led to a substantial reduction in the average alignment length in this data set (364 bp ± 290, compared to the 712 bp ± 520 of the `basidioPAML` data set).

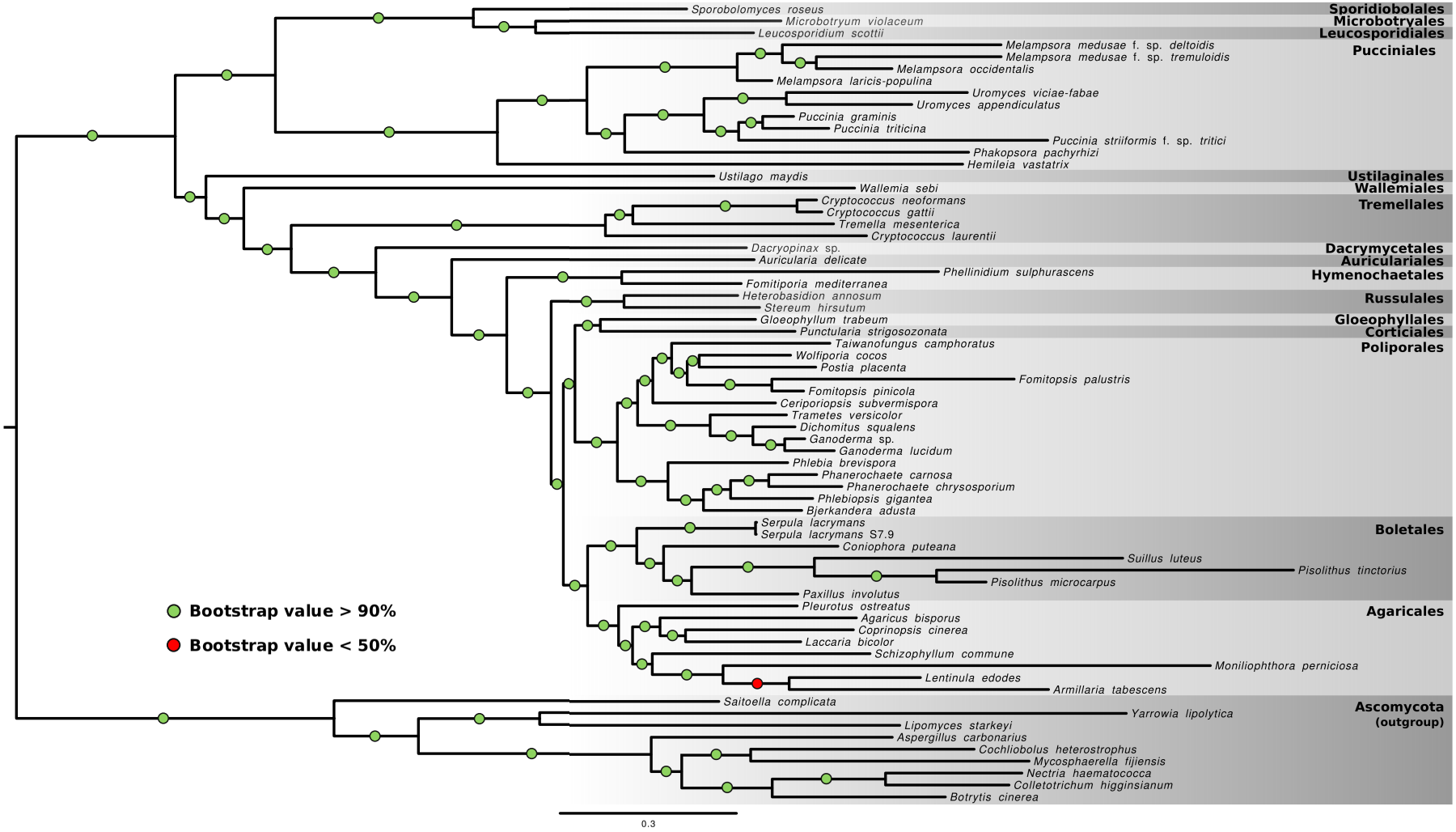

Figure 2.2. Basidiomycota phylogenetic tree. Maximum likelihood tree illustrating the evolutionary relationships among 67 Basidiomycota and Ascomycota species. Names on the right of the figure correspond to the taxonomic order of the respective highlighted taxa. Within the Pucciniales, the three sub-orders defined in Aime (2006) are also presented.

2.4.2 Evolutionary history of the Pucciniales

In a phylogenomic attempt to resolve the phylogenetic backbone of the Pucciniales, Maximum Likelihood (ML) phylogenetic reconstructions were performed. The ML trees obtained among the four data sets were congruent and presented a fully resolved phylogeny with maximum bootstrap support for all branches, except for the position of *Armillaria tabescens* and *Lentinula edodes* in the Agaricales (Figure 2.2). The inclusion of EST data did not have an effect on the general topology of the tree when compared to the trees derived from complete genome data only, despite the substantial increase in missing data. In the Pucciniales, the three previously proposed main sub-orders (Aime, 2006) were recovered with strong support as well as with the same branching order, with *H. vastatrix* diverging early within the rust fungi taxonomic order.

2.4.3 Episodic positive selection at the origin of the rust fungi

Detection of episodic positive selection was undertaken using two data sets aimed at studying the evolutionary origin of the Pucciniales but targeting two different foreground branches. While the `basidioPAML` data set targets the foreground branch leading to the most recent common ancestor of the *Melampsora* and *Puccinia* genera, the `basidioPAML_Hv` data set, which includes EST data from *H. vastatrix*, targets the foreground branch representing the most ancestral split among the Pucciniales. Based on the GUIDANCE filtered alignments, signatures of positive selection were uncovered in 328 genes (33.3% of the data set), after the FDR correction, in the `basidioPAML` data set. In the `basidioPAML_Hv` data set 104 (19.6%) genes were found to be under positive selection for the same FDR threshold. When the same branch-site analysis was performed on GBLOCKS filtered alignments, signatures of positive selection were found in 216 genes (21.9%) for the `basidioPAML` data set and 100 genes (18.8%) for the `basidioPAML_Hv`. Considering both analyses, 177 (82%) positively selected genes were mutually detected in the `BasidioPAML` data set and 78 (78%) in the `BasidioPAML_Hv` data set. Since GUIDANCE was shown to outperform

GBLOCKS as an alignment filtering tool for positive selection detection analyses, particularly due to its ability to recover false negatives (Jordan and Goldman, 2012), only results from GUIDANCE filtered alignments will be further explored.

Since the branch-site model detects episodic selection acting on the amino acid level, information on the number and profile of the selected amino acids can also be obtained to provide deeper insights. In the `basidioPAML` data set, 2 067 sites were found to be under positive selection with a Posterior Probability (PP) above 0.95 across 274 genes (27.8%), while in the `basidioPAML_Hv` data set, 289 selected sites were also detected across 72 genes (13.6%). To further explore the profile of the selected amino acid sites on both data sets, two main site classes were established to assess their potential adaptive role: (i) *Unique*, sites containing a single variant exclusive to the Pucciniales (strict) or rarely found outside the Pucciniales (relaxed); and (ii) *Diversifying*, sites containing multiple variants exclusive to the Pucciniales (strict) or rarely found outside the Pucciniales (relaxed). Therefore, each site class comprises two sub-classes referring to sites sorted in a strict or relaxed (i.e., with a prevalence of the most common amino acid of at least 70%) fashion. The distribution of the number of selected sites per gene and the proportion of sites in each class is summarized in Figure 2.3.

In both data sets, the majority of the positively selected sites could be assigned to either *Unique* or *Diversifying* classes [`basidioPAML`: 1 991 (96%); `basidioPAML_Hv`: 267 (92%)]. In the `basidioPAML` data set, most sites were assigned to the *Unique* class (1 538 sites, 74%, across 259 genes) even though a non-negligible proportion of *Diversifying* sites were uncovered (453 sites, 22%, across 156 genes). Regarding the `basidioPAML_Hv` data set, there was an increase in the proportion of *Diversifying* sites (90 sites, 31%, across 43 genes) but the majority of the selected sites were still placed in the *Unique* class (177 sites, 61%, across 62 genes). Only a small proportion of sites (4% for `basidioPAML` and 8% for `basidioPAML_Hv`) could not be assigned to neither class. Since the same gene can have positively selected sites from both classes, a distribution of the most prevalent site class per gene is presented in Figure 2.4. While *Unique* sites account for the majority of selected

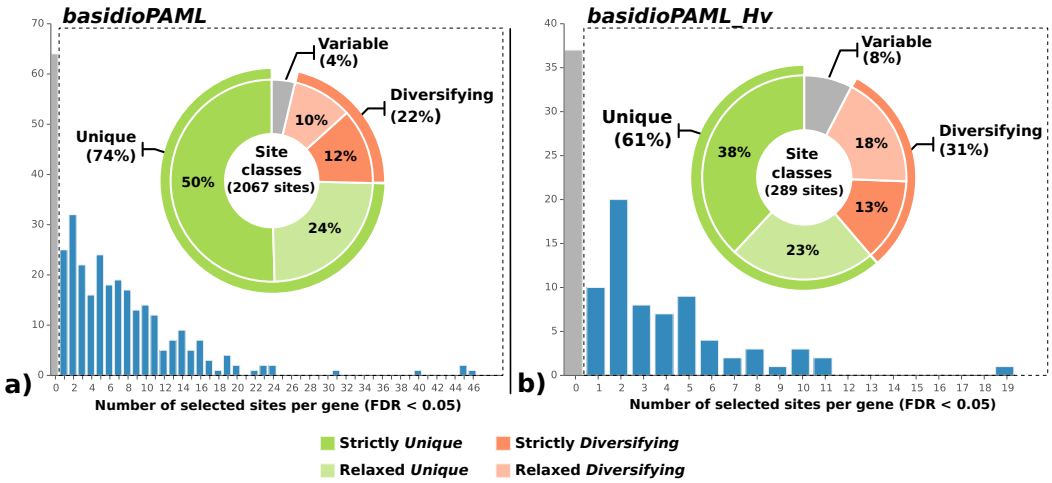

Figure 2.3. Distribution of positively selected sites. Distribution of the number of positively selected amino acid sites after correction of p-values with a false discovery method for (a) the data set containing three rust genomes (`basidioPAML`) and (b) the data set containing the same rust genomes in addition to EST data from *Hemileia vastatrix* (`basidioPAML_Hv`). Embedded in each histogram is a doughnut chart with the distribution of the positively selected sites across the two main site class pairs defined in this study for the `basidioPAML` data set (a) and *basidioPAML_Hv* data set (b). *Unique* sites represent amino acids exclusive and identical in all rust species and *Diversifying* sites represent amino acids exclusive but variable in rust species. The site classes are colour coded with the corresponding legend at the bottom of the figure.

sites in both data sets, the advantage is less pronounced in the `basidioPAML_Hv` data set.

Comparing the `basidioPAML` and `basidioPAML_Hv` data sets, 75 genes had signatures of positive selection in both data sets, while 253 genes were exclusive from the `basidioPAML` data set and 29 were exclusive from the `basidioPAML_Hv` data set. A database containing information about the **PAML** tests for each alignment, including the position and number of selected sites is provided in Table A.1.9.

2.4.4 Estimation of relative evolutionary rates

Substitution rates were estimated using a Bayesian framework for all branches in 614 gene trees represented in most Basidiomycota species in this study. Two trees were excluded from this analysis as they did not include all three Pucciniales species. The mean relative substitution rate of the Pucciniales' root branch was found to be similar to the average substitution rate over all branches and gene alignments

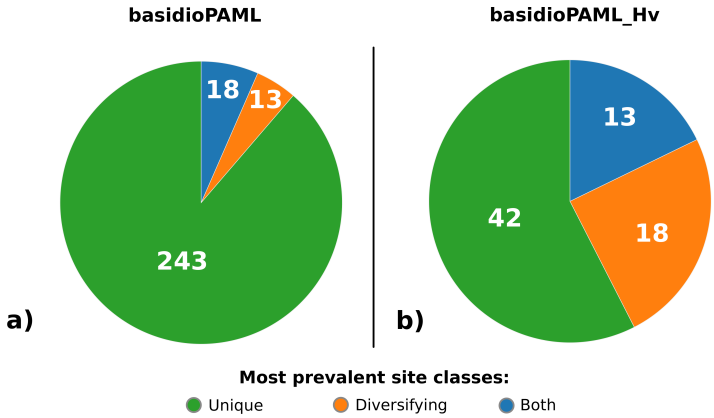

Figure 2.4. Prevalence of positively selected site classes. Pie charts with the distribution of the most prevalent site classes across each positively selected gene for (a) the data set containing only the three rust genomes (`basidioPAML`) and (b) the data set containing the same rust genomes in addition to EST data from *Hemileia vastatrix* (`basidioPAML_Hv`). Site classes are colour coded according to the legend in the right.

(root branch mean = 0.40; all branches mean = 0.43), but the standard deviation found in the root branch was much higher than in the overall substitution rate in all branches (root branch standard deviation = 4.49; all branches standard deviation 0.59). No genes were identified as being evolving at a significantly slower rate in the root branch of the Pucciniales, but 30 genes (5%) were found to be substantially accelerated, compared to the rates of the remaining branches of the corresponding gene. From these genes under accelerated evolution, only 13 were also found to be under positive selection using the branch-site model and these genes were not necessarily the fastest. For example, in the top ten of the fastest evolving genes, only three had signatures of positive selection.

2.4.5 Functional annotation and enrichment

Overall, 357 unique genes (75 shared by both data sets, 29 exclusive from the `basidioPAML_Hv` data set and 253 exclusive from the `basidioPAML` data set) were detected as being under positive selection in either foreground branches. Functional annotation of these 357 genes was obtained using the KOG terminology tables A.1.10 to A.1.11 in page 209. Over 82% (292) of the positively selected genes were classified into 21 specific KOG categories, while 10% (36) had no specific KOG category assigned [“Function unknown” (13) or “General function

prediction only" (23)], and 8% (29) had no hits. Considering the proportion of genes under positive selection against the respective reference of all orthologs analysed, statistical analysis of the under or overrepresentation of the targeted genes for the basidioPAML revealed a relative enrichment (over 1.5 fold change) in genes annotated into "Secondary metabolites biosynthesis, transport and catabolism", "Amino acid transport and metabolism", "Coenzyme transport and metabolism" and "Nuclear structure", and an impoverishment (less than 0.67 fold) in genes annotated into "Chromatin structure and dynamics", "Signal transduction mechanisms" and "Cytoskeleton" (Table A.1.9). However, the Fisher's exact test revealed that only the "Amino acid transport and metabolism" functional class was significantly enriched in positively selected genes. For the basidioPAML_Hv dataset, KOG annotation relative enrichment (over 1.5 fold change) was found in genes as assigned to "Secondary metabolites biosynthesis, transport and catabolism", "Energy production and conversion", "Lipid transport and metabolism" and "Amino acid transport and metabolism", as well as an impoverishment (less than 0.67 fold) in genes annotated as "Inorganic ion transport and metabolism", "Transcription", "RNA processing and modification", "Intracellular trafficking, secretion, and vesicular transport", "Signal transduction mechanisms" and "Cytoskeleton", Figure 2.5. However, no functional classes were significantly over or under represented in this data set according to Fisher's exact test.

To allow a direct comparison between the basidioPAML and basidioPAML_Hv data sets, we performed the same enrichment analysis for the basidioPAML data set with the 531 non-saturated shared genes among data sets as reference in order to account for differences within the same universe of reference genes. Similarly to the respective whole set of orthologs, genes annotated as "Secondary metabolites biosynthesis, transport and catabolism" and "Amino acid transport and metabolism" were enriched as well as genes from the "Carbohydrate transport and metabolism" class. Genes annotated as "Replication, recombination and repair" and "Cytoskeleton" were also found to be impoverished, in addition to "Intracellular trafficking, secretion and versicular transport". Fisher's exact test reported a statistically significant enrichment of the "Amino acid transport and metabolism" functional class as for the complete basidioPAML data set, but further revealed a significant enrich-

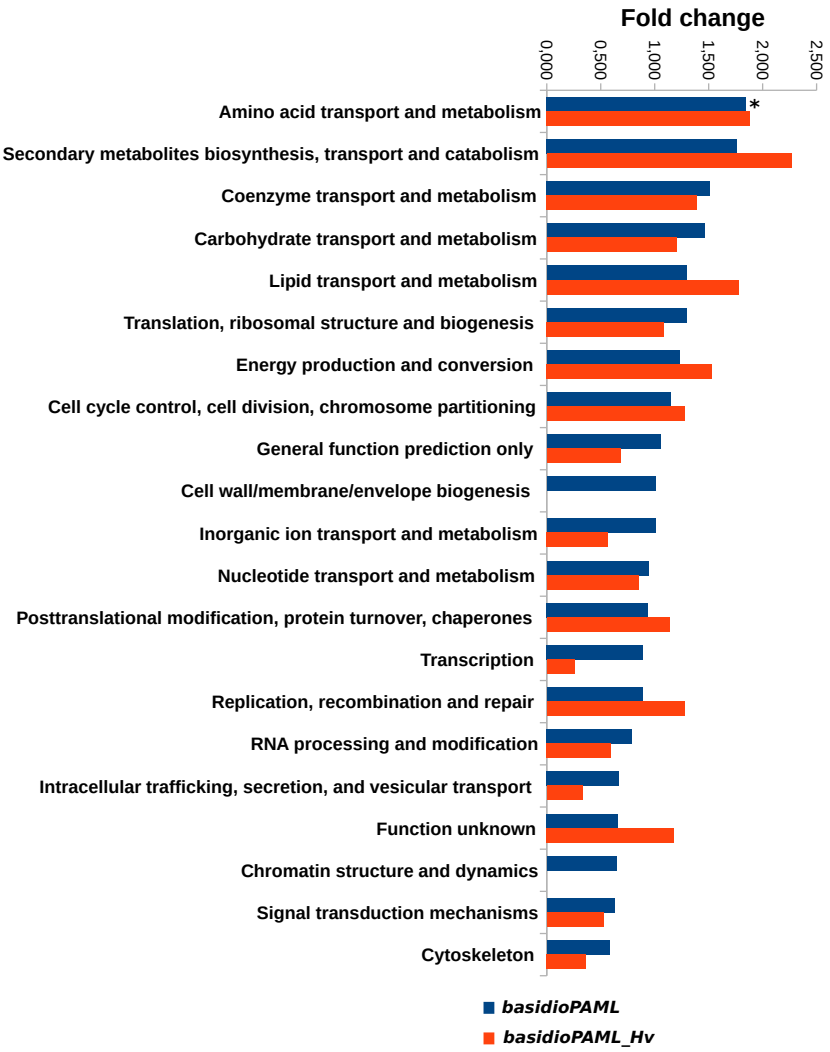

Figure 2.5. Enrichment of functional categories among positively selected genes. Bar chart with the fold change comparing the proportion of genes under positive selection and without positive selection in the y-axis, and the several KOG functional categories in the x-axis. For each functional category, fold change values are presented for the `basidioPAML` and `basidioPAML_Hv` data sets, according to the legend in the top right corner of the figure. Bars with an asterisk ("*") represent statistically significant results for p-value < 0.1.

ment of "Carbohydrate transport and metabolism" and impoverishment of "Signal transduction mechanisms". Among the studied 985 genes, 249 blasted with genes involved in pathogenicity at PHI-base. Out of the 328 positively selected genes in the `basidioPAML`, 93 presented homology with entries in the PHI-base as well as 35 out of the 104 genes under positive selection in the `basidioPAML_Hv`. In both data sets, the Fisher's exact test revealed that positively selected genes were enriched for PHI-base assigned genes putatively involved in pathogenicity, with $p < 0.0001$ for both data sets.

2.5 Discussion

In this work, the role of adaptive genetic variation underlying the origin of the rust fungi (Basidiomycota, Pucciniales) was investigated by developing a phylogenomic framework that assembled and analyzed hundreds of single copy ortholog genes across a broad range of Basidiomycota. A crucial prerequisite for this undertaking is the establishment of a robust phylogeny across the studied Basidiomycota species, particularly of the most basal branches within the Pucciniales. Molecular systematics of the rust fungi has been considerably less explored than for other groups of fungi, but previous studies using a multi-locus approach proposed a division of the order into three sub-orders, Uredinineae, Melampsorineae and Mikronegeriineae (Aime, 2006; Aime et al., 2006; Maier et al., 2007). On par with Ebersberger et al. (2012), our study represents one of the first attempts to use phylogenomics to resolve relationships within the Pucciniales with congruent results, but our data set furthermore includes representatives from all three sub-orders. As previously reported, *H. vastatrix* was placed at the base of the Puccinales in the Mikronegeriineae sub-order, confirming the importance of its inclusion when studying the origin of the rust fungi order (Aime, 2006).

2.5.1 Genome-wide scan for positive selection

After establishing a backbone phylogeny for the studied Basidiomycota, up to 985 non-saturated genes were screened for signatures of positive selection using the branch-site model. This method revealed to be particularly well suited for our investigation, as it is capable of detecting episodes of positive selection specifically on the root branch of the Pucciniales. An analysis pipeline was then constructed to perform a genome scan on two different but related data sets. The `basidioPAML` data set included 985 genes solely retrieved from complete draft genomes but did not include *H. vastatrix*. From those 985 genes, the `basidioPAML_Hv` data set consisted in a sub-sample of 531 genes that could be complemented with EST data from *H. vastatrix*. Therefore, even though our pipeline analyzed the root branch of the Pucciniales in each data set, they represent distinct evolutionary periods, with the

basidioPAML_Hv data set providing the most ancient branch but also significantly less data than the basidioPAML data set. Both data sets presented an impressively high proportion of positively selected genes, suggesting an unexpectedly pervasive role of positive selection shaping the origin of rust fungi: 328 genes (33.3%) showed signatures of positive selection in basidioPAML as compared to 104 genes (19.6%) in basidioPAML_Hv. Even when the same alignments were filtered with GBLOCKS, which is known for its aggressive removal of alignment columns and consequent reduction in power to detected sitewise positive selection, the number of positively selected genes remained high [216 genes (21.9%) for basidioPAML and 100 genes (18.8%) for basidioPAML_Hv]. Studies that employ genome scans to search for positive selection using the branch-site model are numerous but they generally present lower proportions of genes under positive selection. Examples from cetaceans, turtles and fig wasps report proportions of positively selected genes ranging between 4.8-6.2% (Nery et al., 2013; Shaffer et al., 2013; Xiao et al., 2013). Only in a study of humans and chimpanzees was this proportion raised to 16.6%, but this accounts for results across several branches of the phylogenetic tree (Vamathevan et al., 2008). In fungi, the only examples of genome scans aiming at detecting positive selection presented values ranging between 3.2-9% in *Microbotryum* spp. (Aguileta et al., 2010; Aguileta et al., 2012), but in these studies the sites model was used to detect positive selection, which requires the signal to be averaged over all branches (Zhang et al., 2005; Yang, 2007). The discrepancy between our results and those found in the literature may lie in the fact that most of these studies were focused on comparisons between closely related taxa and/or on terminal branches. Indeed, it is well documented that the power of selection detection methods decays quite rapidly when studying closely related species or short branches that have not accumulated sufficient substitutions (Anisimova and Liberles, 2007; Aguileta et al., 2009). In our analyses, positive selection is being searched on a set of non-saturated but substantially divergent sequences and in both data sets the targeted branch is fairly long, which increases the likelihood of finding positive selection. On the other hand, an increase in the number of false positives is also an unavoidable consequence of aligning difficult sets of highly divergent sequences (Jordan and Goldman, 2012), and their occurrence in our data set cannot be ruled out. However, this issue was addressed to the extent that it was possible through the application of several rigor-

ous quality checks and data filters on all alignments. These included the removal of saturated alignments, removal of sites with excessive missing data and alignment filtering using GUIDANCE and GBLOCKS. Given the high number of statistical tests, all original p-values of the branch-site tests were also corrected using a False Discovery Rate method. Therefore, we suggest that the higher proportion of positively selected genes in the root of the Pucciniales may be mainly explained by both a biologically relevant role of positive selection on the origin of this taxonomic group as well as an increase in selection detection power in our data sets.

To further investigate how positive selection might have acted during the early evolution of rust fungi, we took advantage of the ability of the branch-site model to detect selection on specific amino acids on the branch of interest. For the `basidioPAML` data set, 2 067 sites were detected to be under positive selection across 274 genes (27.8%), as compared to the 289 sites across 72 genes (13.6%) on the `basidioPAML_Hv` data set. While the lower proportion of positively selected genes and sites in the latter data set may be attributed to biologically relevant differences that arise from studying different foreground branches, this may also be a reflection of differences in the composition of the data sets, such as the average gene length reduction in the data set containing *H. vastatrix*, as regions with potentially selected sites may not be present in the sequences of this species.

We then explored the profile of the selected amino acids by sorting them among two main site classes: *Unique* and *Diversifying*. By establishing these classes, the aim was to shed light on the adaptive role of the selected sites and how they may have contributed to the evolution of rust fungi. *Unique* and *Diversifying* positively selected sites are potential prime candidates for adaptive drivers underlying the initial divergence and adaptation of rust fungi, but their distinction may be key to distinguish between adaptations that are important for all rust species and remain unchanged throughout their evolution and adaptations that later diversified in different terminal branches of the Pucciniales. Most of the positively selected sites showed a *Unique* class pattern in both data sets, though its proportion decreases in the `basidioPAML_Hv` data set, with the inclusion of the ancient *H. vastatrix* branch, which is also reflected on the prevalence of each site class in the positively selected

genes. This result is expected because the basidioPAML_Hv data set includes more divergent species, but it also stresses the importance and impact of including an ancient lineage when studying the evolutionary history of a taxonomic group. In either case, this seems to reveal that adaptive changes that occur at the root of the Pucciniales are more likely to remain conserved across all rust species possibly in order to conserve a newly acquired adaptive trait. On the other hand, some positively selected sites were subject of further modification since the origin of rust fungi and may be responsible for species or genus specific alterations, although it cannot be concluded whether such modification was due to a relaxation of selective pressure or continuous action of positive selection. Nevertheless, our results show that 96% and 92% of the positively selected genes in basidioPAML and basidioPAML_Hv data sets, respectively, contain both *Unique* and *Diversifying* sites, suggesting that natural selection may shape the same gene in different ways with potentially different outcomes during the evolution of a species group. This new layer of information provided by modern methods for the detection of positive selection at the macro-evolutionary scale may allow researchers to move beyond the simple identification of genes under positive selection and grant them the power to detect the specific sites targeted by natural selection and how their variants evolve henceforth. Indeed, this information should be valuable for future studies investigating the functional aspects of the positively selected genes detected in this study for rust fungi, since they will provide the most likely sites responsible for functional change.

2.5.2 Assessing the evolutionary rate on the origin of Pucciniales

To further investigate the evolutionary rate on the basal branch of the Pucciniales, we assessed the substitution rate on that specific branch across the 612 genes with the highest taxa representation in our data set, including all three Pucciniales species with complete genomes. *Hemileia vastatrix* was not included because the fragmented and overall lower quality nature of the EST data could bias estimates of substitution rates and produce misleading results. On average, the molecular rate of the root branch of the Pucciniales seems to be evolving on par with the average

molecular rate of all other branches in the tree, but a significant heterogeneity in the substitution rates was also present, with a strong bias towards fast evolving genes since all outlier genes were evolving more rapidly. This stems from the fact that a single point substitution rate per gene is being compared to an average over all substitution rates per gene that tends to homogenize a sample. Thirty genes (5%) were found to be significantly accelerated but only 13 were also found to be under positive selection. While it is easy to conceive that a gene found to be under positive selection may not be under an accelerated substitution rate, particularly if positive selection acts upon a limited number of sites, the explanation for the absence of positive selection on genes under accelerated evolution at the amino acid level is less straightforward. However, two explanations seem plausible. One can be attributed to the assembly of the data sets themselves. Since tests for positive selection are more sensitive to misaligned and ambiguous alignment columns, these were filtered with GUIDANCE for the SLIMCODEML analyses but not for the MCMCTREE analyses. Therefore, alignment regions where selection could actually be occurring may be visible in the evolutionary rates analysis, but absent in the branch-site test. By repeating the branch-site test for data sets without GUIDANCE filtering the number of overlapping genes between analyses raised to 18 (data not shown), suggesting that the signature of positive selection could be indeed present in the complete alignments of some genes under accelerated evolution. The second possible explanation can be attributed to a relaxation of purifying selection on these genes, rather than the action of positive selection. In a case of relaxed purifying selection, the rate of nonsynonymous mutations does not overly exceed the rate of synonymous mutations and therefore, there is no signature of positive selection. Indeed, individual inspection of branch-site test results for the non-overlapping genes revealed that the rate of synonymous substitution was high relatively to non-synonymous substitutions, supporting the scenario of relaxed purifying selection. In such case, the relaxation of purifying selection would allow for a greater accumulation of both synonymous and non-synonymous substitutions, which would produce the observed results. Such genes could be under a relaxation of purifying selection due to temporary loss of function (Wang et al., 2004), subfunctionalization associated with gene duplication (Wei and Ge, 2011), or as precursors of phenotypic plasticity (Hunt et al., 2011). Even though these genes lack the intriguing signature of positive selection, they should

warrant further investigation as to why purifying selection was suddenly relaxed on genes that are conserved across a broad range of Basidiomycota species.

2.5.3 Functional enrichment analyses

The impact of studying different foreground branches that span different time periods between the `basidioPAML` and `basidioPAML_Hv` data sets was further explored by investigating the enrichment of functional classes for the positively selected genes in each data set. Given the substantial overlap of positively selected genes between data sets (75 genes), the enriched functional classes were also found to be similar. The metabolic related classes of "Amino acid transport and metabolism" and "Secondary metabolites biosynthesis, transport and catabolism" were found to be the most enriched for positively selected genes in both data sets, with the first mentioned class showing statistical significance in the `basidioPAML` data set. When comparing directly these data sets at the level of the 531 shared genes, we found that the "Carbohydrate transport and metabolism" class was also significantly enriched for the `basidioPAML` data set. Therefore, enrichment of functional classes related to the metabolism of amino acids, carbohydrates and secondary metabolites seems to be a common trend across both data sets and is likely to reflect a key process of adaptation that rusts in general faced during their early evolution. We further note that the positively selected genes assigned to these enriched classes were very similar between data sets. Nevertheless, differences were also found for functional classes such as "Lipid transport and metabolism" and "Energy production and conversion", both of which presented higher fold changes in the `basidioPAML_Hv` data set. These differences may be a reflection of natural selection targeting slightly different functional classes during the early evolution of all Pucciniales.

In such a focused group of plant pathogens it is not surprising to find that positive selection may have favoured genes directly related with pathogenicity. In fact, the positively selected genes in both data sets were enriched for genes with PHI-base assignment, which have been previously demonstrated to be involved in plant-pathogen interactions and to have known roles in the infection process of several fungal species. Even though the pathogenicity related genes present in PHI-base

have been mostly characterized in different fungal species, which makes extrapolations less straightforward, this provides a potential link between our positively selected genes and their contribution to the pathogenic process. For instance, genes shown to produce a mutant phenotype of loss of pathogenicity were found to be under positive selection in both evolutionary times, such as *BasidioOnly3480* gene, encoding a UDP-glucose dehydrogenase (EC:1.1.1.22), ortholog to *ugd1* gene required for the pathogenicity of *Cryptococcus neoformans* (Bar-Peled et al., 2004); only in the most ancestral branch, such as gene *BasidioOnly3181*, encoding a 3-isopropylmalate dehydrogenase, ortholog to *leu2* gene that reduces virulence in *Saccharomyces cerevisiae* (Goldstein and McCusker, 2001); and only in the most recent foreground branch studied, such as gene *BasidioOnly3451*, encoding a mitochondrial homoaconitate hydratase, ortholog to the *Aspergillus fumigatus* gene *lysF* required for pathogenicity (Liebmann et al., 2004). Moreover, among the seven selected genes comprised in the enriched category "Secondary metabolites biosynthesis, transport and catabolism" in both data sets, five are referenced as associated with loss of pathogenicity or reduced virulence in *Magnaporthe grisea*, *Stagonospora nodorum* and *Colletotrichum lagenarium* (Perpetua et al., 1996; Sun et al., 2006). These include gene *BasidioOnly2693*, encoding an ATP-binding cassette transporter, ortholog to the multidrug resistance efflux pump ABC3 gene of *Magnaporthe grisea*, which is required for host penetration and for survival during oxidative stress mounted by the host through the efflux of toxic compounds (Sun et al., 2006), or *BasidioOnly3214*, ortholog to THR1 reductase gene of *Colletotrichum lagenarium*, essential for the appressorium melanization process, which is a requirement for successful host penetration (Perpetua et al., 1996). Interestingly, rust fungi and other biotrophs have suffered a loss of secondary metabolism genes, possibly due to the reduction in the need of degrading plant cell wall biomass and availability of amino acids from the living host (Spanu, 2012). Moreover, since these genes may also be capable of triggering an effective host defense their removal may reduce the opportunity to elicit a rejection from the host (Spanu, 2012). It is thus possible that the action of positive selection on these genes is another way for the pathogen to overcome the host defenses by introducing variation that avoids recognition by the host.

Positively selected genes were also found to be enriched in genes involved in "Amino acid transport and metabolism", encoding enzymes with crucial roles which catalyze important steps or regulate several biosynthetic pathways. Most of the genes assigned to this functional category showed a signature of positive selection only in the foreground branch leading to the most recent common ancestor of the *Melampsora* and *Puccinia* genera (`basidioPAML`). These include genes codifying for key step enzymes, such as a putative ATP phosphoribosyltransferase (*BasidioOnly3456*), catalyzing the first step and controlling the rate of histidine biosynthesis (`EC 2.4.2.17`).

Interestingly, genes encoding two of the three key enzymes governing polyamine metabolism are represented in this list [*BasidioOnly3343* predicted ornithine decarboxylase (ODC) (`E.C.4.1.1.17`) and *BasidioOnly3535* predicted S-adenosylmethionine decarboxylase (SAMDC) (`E.C.4.1.1.50`)] (Valdés-Santiago et al., 2012b). Moreover, a gene (*BasidioOnly3294*) annotated as a predicted arginase, which converts arginine into ornithine, the substrate of ODC, was also found to be under positive selection but only in the most ancestral foreground branch. The two decarboxylases are the rate-limiting enzymes of polyamine biosynthesis playing a central role on the fine tune regulation mechanism controlling intracellular polyamines pools. It has been demonstrated that during host-fungus interaction, polyamine metabolism suffers striking changes in response to infection (Valdés-Santiago et al., 2012b). Polyamines are essential for growth and have been implicated in the regulation of both cell proliferation and differentiation processes, such as dimorphism, spore germination, appressorium formation and conidiation (Ruiz-Herrera, 1994). By modulating development and differentiation, in some way or another, polyamines regulate the virulence of animal and plant fungal pathogens. Bailey et al. (2000) have shown that the ODC gene is essential in *Septoria nodorum* to obtain polyamines from the host plant for normal growth during infection. Additionally, *Ustilago maydis* mutants affected in the SAMDC gene were shown to be completely avirulent to maize (Valdés-Santiago et al., 2012a). An enrichment in positively selected genes involved in amino acid metabolism is consistent with a high primary metabolism activity observed in the invading rust fungi (Duplessis et al., 2011), since acquisition

of nutrients for the development of haustoria within the host plant is crucial to the success of rust pathogen biotrophic interactions.

Positively selected genes belonging to the "Carbohydrate transportation and metabolism" class were also found to have an impact on pathogenicity of other fungi. The *BasidioOnly3559* gene, ortholog to *tps1*, encodes for a synthase subunit of the trehalose-6-P synthase/phosphatase complex which synthesizes the storage carbohydrate trehalose. It was shown in *Magnaporthe oryzae*, that *tps1* has regulatory functions that control the expression of virulence associated genes and plays a pivotal role in the establishment of plant disease (Wilson et al., 2007). The *BasidioOnly4049* gene encodes for a serine/threonine protein phosphatase, ortholog to *Saccharomyces cerevisiae sit4* gene, which is important for hyphal growth and virulence in *S. cerevisiae* through the regulation of cell wall biogenesis, osmosensing and protein translation (Lee et al., 2004). The *ugd1* gene (*BasidioOnly3480* predicted UDP-glucose dehydrogenase) is required for the metabolism of UDP-glucuronic acid, which in turn is essential for the capsule formation and pathogenicity of *Cryptococcus neformans* (Griffith et al., 2004). In general, genes involved in "Carbohydrate transportation and metabolism" are important since very early stages of fungal pathogenicity, including the mobilization of nutrients from spores for fungal development, but also for melanin biosysthesis in appressoria formation, lytic activity and camouflage. For instance, the activity of chitin deacetylases, with convert fungal chitin to chitosan, avoiding recognition by host chitinases, enables fungal camouflage and contributes to pathogenicity (Talhinhas et al., 2014).

Overall, these results reveal the usefulness and predictive power of detecting positive selection over conserved genes that apparently have a significant role in the evolution of rusts in general and may have been involved in their origin. Other genomic features and patterns unique to the rust fungi and associated with their origin, such as gene duplication/loss patterns, were also recently identified (Pendleton et al., 2014). In their study, Pendleton et al. (2014) revealed that the origin of the rust fungi was associated with the loss of a substantial number of genes (1217) and a relatively small number of duplications (248). This is also consistent with our observation that the rust species included in our work shared less orthologs with

other Basidiomycetes than any other Basidiomycota species between each other. However, despite the considerable gene family losses and contractions at the origin of the group, the three individual rust species included in Pendleton et al. (2014) (*Cronartium quercuum* f. sp. *fusiforme*, *Melampsora lirici-populina* and *Puccinia graminis* f. sp. *tritici*) exhibited disproportionally high amounts of species-specific gene duplications. Considering these results along with the pervasive signal of positive selection found in our study, this suggests that the transition of the rust fungi to obligate biotrophy required major changes in both the composition of gene families and in the nucleotide sequences of conserved genes, followed by rapid species specific changes that would enable adaptation to their hosts. In the future, it will be interesting to investigate if similar patterns underlie the origin of other phylogenetically unrelated obligate biotrophs. Our framework also provided a useful and comprehensive hypothesis generator data for future studies focusing on the functional impact of the adaptive changes found in these genes. Indeed, further detailed investigation of the genes targeted by natural selection on the origin of rust fungi, particularly those related to metabolism, may yield new insights on the precursors of the main features of rust fungi, and these insights can be compared and verified in other biotrophic taxa. Additionally, the site classes for the positively selected amino acids defined in our study provide an additional and interesting layer of information that will require attention when investigating exactly how changes in the nucleotide sequences translate into specific adaptations and impact the fitness of individuals. It will be interesting to see if the proportions of *Unique* and *Diversifying* site classes are similar across a wide range of taxa. Despite the large number of genes, our study has detected positive selection on a single main branch of the Pucciniales. As more rust genomes get sequenced, it will become possible to investigate the action of positive selection not only on single branches but, more excitingly, on all branches since their origin up to the most terminal taxa in order to see how natural selection is shaping genes along the entire evolutionary history of rust fungi.

2.6 Supplementary information

2.6.1 Positive selection on conserved amino acid sites

During the development of the work described above, an unexpected signal of positive selection was found on amino acids that are conserved across all or most of the studied taxa in this study. In the following section, we present a first assessment of the results to further explore the role of positive selection acting on these amino acids, which were termed *Conserved* sites.

2.6.1.1. Results

The present results concern to the 216 and 100 positively selected genes detected for the `basidioPAML` and `basidioPAML_Hv` data sets, respectively, using the Gblocks alignment filtering and including the FDR correction for multiple testing. In addition to the *Unique* and *Diversifying* site classes previously established, we defined an additional class named *Conserved*, which consists of sites containing a single amino acid across all species (strict), or the most common variant found in more than 70% of the studied species (relaxed). The distribution of the number of positively selected sites per gene and the proportion of sites in each of the three classes is summarized in Figure 2.6. While the majority of the positively selected sites remained assigned to the *Unique* and *Diversifying* classes, a non-negligible proportion of *Conserved* sites under positive selection was uncovered (`basidioPAML`: 196 sites, 15% across 113 genes; `basidioPAML_Hv`: 77 sites, 24% across 45 genes). Concerning the distribution of the most prevalent site class per gene (Figure 2.7), the majority of the positively selected genes in the `basidioPAML` data set contained predominantly *Unique* sites followed by a lower number of conserved sites. However, this trend is reversed in the `basidioPAML_Hv`, in which *Conserved* sites are predominant in the majority of the positively selected genes.

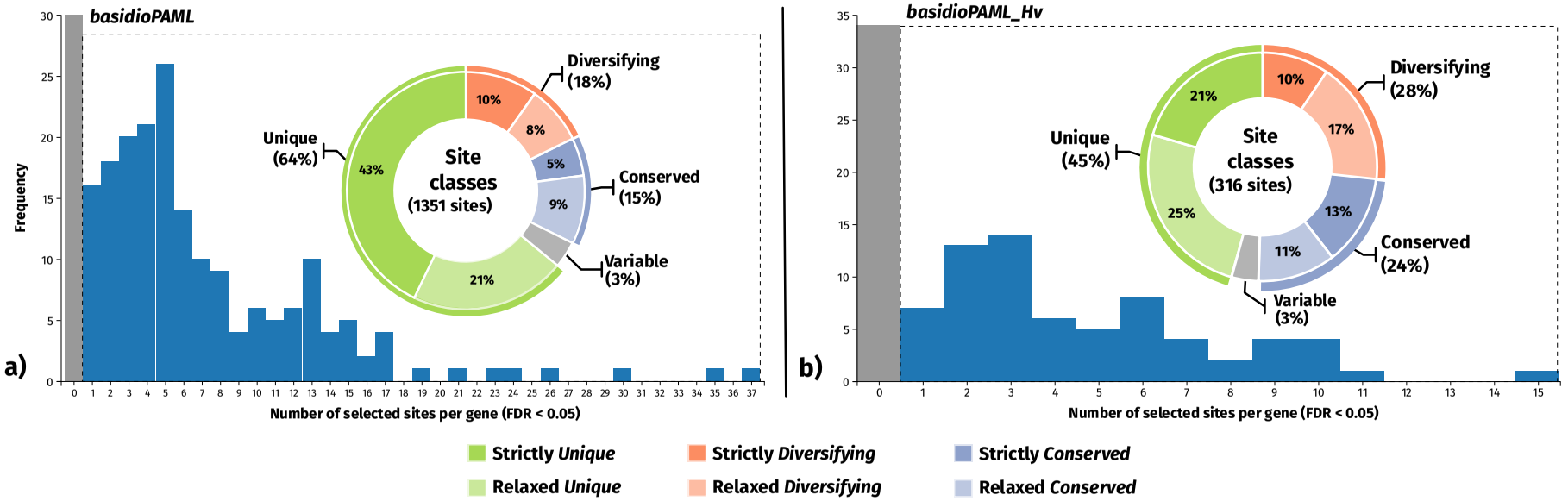

Figure 2.6. Distribution of the number of positively selected amino acid sites after correction of p-values with a false discovery method for (a) the data set containing three rust genomes (`basidioPAML`) and (b) the data set containing the same rust genomes in addition to EST data from *Hemileia vastatrix* (`basidioPAML_Hv`). Embedded in each histogram is a doughnut chart with the distribution of the positively selected sites across the three site class pairs defined in this study for (a) the `basidioPAML` data set and (b) `basidioPAML_Hv` data set. *Unique* sites represent amino acids exclusive and identical in all rust species, *Diversifying* sites represent amino acids exclusive but variant in rust species and *Conversed* sites represent conserved amino acids across all species. The site classes are colour coded with the corresponding legend on the bottom of the figure.

The identification of positive selection on *Conserved* sites occurred in amino acids that have a broad range of encoding codons. In some cases, shifting from one such codon to another may require a non-synonymous mutation, as in the transition from TCG to AGC, both of which encode a Serine residue.

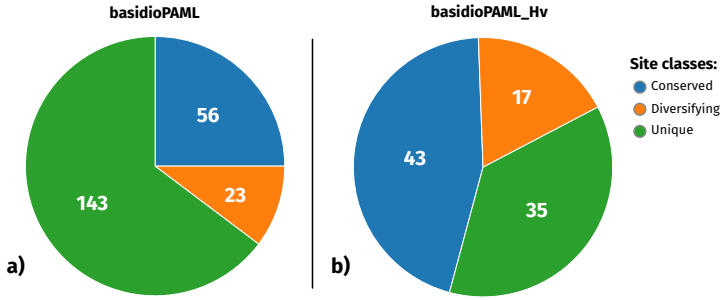

Figure 2.7. Pie charts with the distribution of the most prevalent site classes across each positively selected gene for the data set containing (a) only the three rust genomes (`basidioPAML`) and (b) the data set containing the same rust genomes in addition to EST data from *Hemileia vastatrix* (`basidioPAML_Hv`). Site classes are colour coded according to the legend in the right.

Such kind of shifts in codon usage on the root branch of the Pucciniales would produce an excess of non-synonymous substitutions and generate a signal of positive selection, even though the final amino acid is the same. Differences in preferred and non preferred codon between rusts and non rusts for the *Conserved* amino acid sites were assessed for each data set and the results are summarized in Figure 2.8.

In both data sets, all strictly *Conserved* and most relaxed *Conserved* sites are found in codons encoding a Serine residue, where a substantial shift in codon preference was detected. While there was a clear preference for AGY codons in non rust fungi (73-90%), this preference was diluted in the rust fungi (36-42%) in favor of TCN codons. Besides Serine residues, relaxed *Conserved* sites were also found in encoding Arginine and Leucide residues, in the `basidioPAML` and `basidioPAML_Hv` data sets, respectively, where shifts in codon preferences were also observed. In both cases, the preferred codon in non rust fungi is less prevalent in rust fungi.

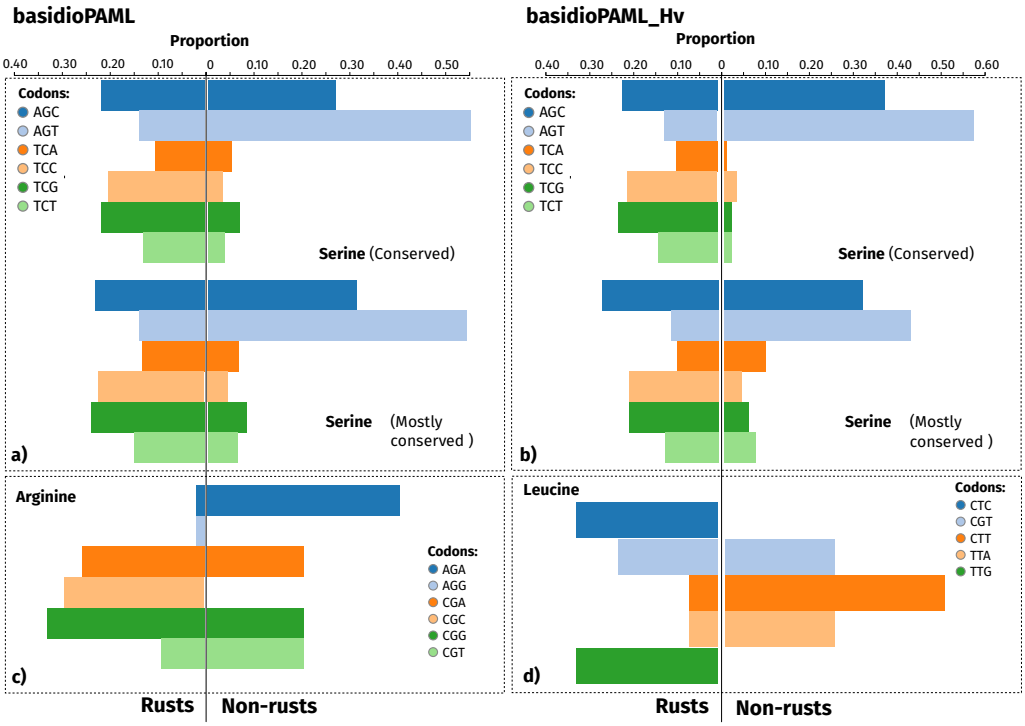

Figure 2.8. Bar charts representing the proportion of codon usage for different amino acid residues and data sets between rust and non-rust species. Plots (a) and (c) refer to the data set containing only three rust genomes (`basidioPAML`), while plots (b) and (d) refer to the data set containing the same three rust genomes in addition to *Hemileia vastatrix* (`basidioPAML_Hv`). For the serine residue, two plots are shown corresponding to sites strictly conserved (all species possess the same amino acid) and mostly conserved (at least 70% of the species possess the same amino acid). Codons are colour coded according to the legend next to each plot.

2.6.1.2. Discussion

The occurrence of positive selection on conserved amino acids is an intriguing result because the ancestral and derived amino acids are the same. This signal of positive selection on conserved sites was, nevertheless, detected on a high proportion of conserved sites whenever there was a codon shift for the same amino acid, which is possible for amino acids such as leucine, arginine and serine. For example, the positively selected codon 103 in the alignment "*BasidioOnly3490*", which encodes a serine residue for all studied species, is by an AGY type codon in rust fungi and a TCN type codon in non rust fungi. This suggests that the TCN type codon is the ancestral state for the serine residue and that at some point in the early evolution of rust fungi, there was a transition to an AGY type codon that is shared among all studied rust fungi. The transition between these codon types would require at least

two non-synonymous substitutions, if the ancestral codon was a TCY codon type, or three non-synonymous substitutions if the ancestral codon was a TCR codon type. This codon change would effectively produce an excess of non-synonymous substitutions relatively to synonymous substitutions, which would be interpreted by the software as a signal of positive selection. Moreover, among the conserved sites, different codons were predominantly found when comparing rust fungi to other basidiomycetes. In serine residues, there was a loss of preference for certain codons that were mostly found on other basidiomycetes, and in arginine and leucine residues, there was a marked shift in preference for different codons. The codon bias phenomenon is known in other species, such as Saccharomyces cerevisiae and Escherichia coli (Tuller et al., 2010), but whether it is driven by natural selection or not remains an open question. Although **PAML** flagged theses sites as being under positive selection, this molecular signature may not be necessarily generated only by this process. Thus, four hypotheses to explain this result are presented: Positive selection, strong purifying/stabilizing selection with or without simultaneous double-nucleotide substitution and mutational bias.

If this codon shift was indeed due to sequential non-synonymous mutations due to the action of natural selection, it implies that these highly conserved sites suffered a non-synonymous change into a different amino acid, only to return to the original amino acid using a different codon that conferred some advantage. The role of natural selection on codon usage has been much debated and little is still known about how it operates and impacts on the phenotype and fitness of individuals but it has been proposed that it may have a deep impact on the efficiency and accuracy of protein expression as well as on the speed of translation elongation (Neafsey and Galagan, 2007; Hershberg and Petrov, 2008). Indeed, even small changes in the usage of codons were demonstrated to affect the expression levels of proteins with adaptive consequences, as exemplified by the *adh* gene and the ability of the Drosophila flies carrying an over-expressed *adh* gene derived from a codon change to tolerate ethanol (Carlini, 2004). Another interesting discovery that addresses codon choice is the fact that some codons, named “duons”, may encode two types of information (Stergachis et al., 2013). One is interpreted by the genetic code to assemble proteins, and the other is actually a transcription factor-binding regulatory

code that influences gene expression. Indeed, these transcription biding sites can exert a considerable pressure in codon choice in order to maintain the association (Stergachis et al., 2013). Under this scenario, the original codon may represent a local maximum in the fitness landscape. If the intermediary amino acid is neutral or slightly deleterious, it is conceivable that this may present an opportunity for the organism to explore a different codon for the same amino acid that represents a higher fitness maximum. In this way, natural selection could select specific codons that would provide an advantage during the early evolution of the rust fungi.

However, if both ancestral and derived variants have similar fitness effects on the organism this could represent neutral stochastic variation that maintained the same amino acid due to the action of purifying/stabilizing selection. It is also possible that the two or three non-synonymous mutations required to change the codon may occur simultaneously, therefore bypassing the transient change in amino acids. Even though the probability of these mutations occurring in a single generation is incredibly low, there is evidence that simultaneous double nucleotide substitutions may have a higher frequency than expected, particularly in serine residues (Averof, 2000), and several mechanisms have been proposed on how this may occur (Golding and Glickman, 1985; Hampsey et al., 1988). In each case, these would present neutral changes with no fitness effect for the organism.

Another possibility is the occurrence of a mutational bias, which is particularly prevalent on high recombinant regions and is though to be a consequence of the recombination process itself (Marais et al., 2001). It has been shown that a bias in codon usage may reflect a variation of mutational patterns with recombination rather than the effect of selection for certain codons, because the codon usage bias is present in both coding and non-coding regions (Marais et al., 2001).

Although we performed the **PAML** analyses in a similar fashion as numerous previous studies, our results represent the first description of this evolutionary phenomenon within a framework of a genome scan for positive selection, as far as we know. Therefore, without a comparison basis from other genome scans for positive selection on conserved amino acids, it is not possible to draw an integrative picture on how

the proportion of *Conserved* sites fares in different taxa and branches. Further studies investigating the functional consequences of different codons encoding the identified *Conserved* amino acid at the protein level will be paramount to better understand the relative contributions of selective and neutral mechanisms and to assess their interplay in generating these molecular signatures at the macro-evolutionary scale. If these results stem from the action of natural selection, this may reveal a much more preponderant and adaptive role of codon choice on the evolution of rust fungi. Alternatively, if they arise from neutral processes that generate a similar molecular signature to that of positive selection, future studies employing genome-wide scans for positive selection using d_N/d_S methods should account for this potentially confounding effect when interpreting the results.

2.7 References

Abascal, F., R. Zardoya, and M. J. Telford (2010). „TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations.“ *Nucleic Acids Research* 38, W7–13.

Aguileta, G, J Lengelle, S Marthey, H Chiapello, F Rodolphe, A Gendrault, R Yockteng, E Vercken, B Devier, M. C. Fontaine, et al. (2010). „Finding candidate genes under positive selection in Non-model species: examples of genes involved in host specialization in pathogens.“ *Molecular Ecology* 19.2, pp. 292–306.

Aguileta, G., G. Refrégier, R. Yockteng, E. Fournier, and T. Giraud (2009). „Rapidly evolving genes in pathogens: methods for detecting positive selection and examples among fungi, bacteria, viruses and protists.“ *Infection, Genetics and Evolution* 9.4, pp. 656–670.

Aguileta, G., J. Lengelle, H. Chiapello, T. Giraud, M. Viaud, E. Fournier, F. Rodolphe, S. Marthey, A. Ducasse, A. Gendrault, et al. (2012). „Genes under positive selection in a model plant pathogenic fungus, Botrytis.“ *Infection, Genetics and Evolution* 12.5, pp. 987–996.

Aime, M. C. (2006). „Toward resolving family-level relationships in rust fungi (Uredinales)“. *Mycoscience* 47.3, pp. 112–122.

Aime, M. C., P. B. Matheny, D. A. Henk, E. M. Frieders, R. H. Nilsson, M. Piepenbring, D. J. McLaughlin, L. J. Szabo, D. Begerow, J. P. Sampaio, et al. (2006). „An overview of the higher level classification of Pucciniomycotina based on combined analyses of nuclear large and small subunit rDNA sequences.“ *Mycologia* 98.6, pp. 896–905.

Anisimova, M and D. A. Liberles (2007). „The quest for natural selection in the age of comparative genomics.“ *Heredity* 99, pp. 567–579.

Averof, M. (2000). „Evidence for a High Frequency of Simultaneous Double-Nucleotide Substitutions“. *Science* 287.5456, pp. 1283–1286.

Bailey, A, E Mueller, and P Bowyer (2000). „Ornithine Decarboxylase of Stagonospora (Septoria) nodorum Is Required for Virulence toward Wheat“. *The Journal of Biological Chemistry* 275.19, pp. 14242–14247.

Bar-Peled, M., C. L. Griffith, J. J. Ory, and T. L. Doering (2004). „Biosynthesis of UDP-GlcA, a key metabolite for capsular polysaccharide synthesis in the pathogenic fungus Cryptococcus neoformans“. *Biochemical Journal* 381.1, pp. 131–136.

Barrett, R. and H. Hoekstra (2011). „Molecular spandrels: tests of adaptation at the genetic level“. *Nature Review Genetics* 12.11, pp. 767–780.

Barrett, R. D. H. and D. Schluter (2008). „Adaptation from standing genetic variation“. *Trends in Ecology and Evolution* 23.1, pp. 38–44.

Benner, S. A., S. O. Sassi, and E. A. Gaucher (2007). „Molecular paleoscience: systems biology from the past.“ *Advances in enzymology and related areas of molecular biology* 75, pp. 1–132.

Cantu, D., V. Segovia, D. MacLean, R. Bayles, X. Chen, S. Kamoun, J. Dubcovsky, D. G. O. Saunders, and C. Uauy (2013). „Genome analyses of the wheat yellow (stripe) rust pathogen Puccinia striiformis f. sp. tritici reveal polymorphic and haustorial expressed secreted proteins as candidate effectors.“ *BMC Genomics* 14, p. 270.

Carlini, D. B. (2004). „Experimental reduction of codon bias in the Drosophila alcohol dehydrogenase gene results in decreased ethanol tolerance of adult flies.“ *Journal of Evolutionary Biology* 17.4, pp. 779–785.

Carroll, S. B. (2008). „Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution.“ *Cell* 134.1, pp. 25–36.

Chen, F., A. J. Mackey, J. K. Vermunt, and D. S. Roos (2007). „Assessing performance of orthology detection strategies applied to eukaryotic genomes.“ *PloS ONE* 2.4, e383.

Dongen, S (2000). „Graph clustering by Flow Simulation“. PhD thesis. University of Utrecht.

Duplessis, S., C. A. Cuomo, Y.-C. Lin, A. Aerts, E. Tisserant, C. Veneault-Fourrey, D. L. Joly, S. Hacquard, J. Amselem, B. L. Cantarel, et al. (2011). „Obligate biotrophy features unraveled by the genomic analysis of rust fungi.“ *Proceedings of the National Academy of Sciences of the United States of America* 108.22, pp. 9166–9171.

Ebersberger, I., S. Strauss, and A. von Haeseler (2009). „HaMStR: Profile hidden markov model based search for orthologs in ESTs“. *BMC Evolutionary Biology* 9.1, p. 157.

Ebersberger, I., R. de Matos Simoes, A. Kupczok, M. Gube, E. Kothe, K. Voigt, and A. von Haeseler (2012). „A consistent phylogenetic backbone for the fungi.“ *Molecular Biology and Evolution* 29.5, pp. 1319–1334.

Eddy, S. R. (2010). „HMMER User ' s Guide“. *HMMER MANUAL* March, pp. 0–93.

Enright, a. J., S Van Dongen, and C. a. Ouzounis (2002). „An efficient algorithm for large-scale detection of protein families.“ *Nucleic Acids Research* 30.7, pp. 1575–1584.

Fernandez, D., P. Talhinhas, and S. Duplessis (2013). „Rust Fungi: Achievements and Future Challenges on Genomics and Host–Parasite Interactions“. In: *Agricultural Applications*. Vol. 11, pp. 315–341.

Golding, B. G. and B. W. Glickman (1985). „Sequence-directed mutagenesis: Evidence from a phylogenetic history of human alpha-interferon genes“. *Proceedings of the National Academy of Sciences of the United States of America* 82, pp. 8577–8581.

Goldstein, A. L. and J. H. McCusker (2001). „Development of Saccharomyces cerevisiae as a model pathogen: A system for the genetic identification of gene products required for survival in the mammalian host environment“. *Genetics* 159.2, pp. 499–513.

Griffith, C. L., J. S. Klutts, L. Zhang, S. B. Levery, and T. L. Doering (2004). „UDP-glucose dehydrogenase plays multiple roles in the biology of the pathogenic fungus Cryptococcus neoformans“. *Journal of Biological Chemistry* 279.49, pp. 51669–51676.

Hampsey, M., J. Ernst, J. Stewart, and F Sherman (1988). „Multiple Base-pair Mutations in Yeast“. *Journal of Molecular Biology* 201, pp. 471–486.

Hershberg, R. and D. a. Petrov (2008). „Selection on Codon Bias“. *Annual Review of Genetics* 42.1, pp. 287–299.

Hunt, B. G., L. Ometto, Y. Wurm, D. Shoemaker, S. V. Yi, L. Keller, and M. a. D. Goodisman (2011). „Relaxed selection is a precursor to the evolution of phenotypic plasticity.“ *Proceedings of the National Academy of Sciences of the United States of America* 108.38, pp. 15936–15941.

Jordan, G. and N. Goldman (2012). „The effects of alignment error and alignment filtering on the sitewise detection of positive selection“. *Molecular Biology and Evolution* 29.4, pp. 1125–1139.

Katoh, K. and D. M. Standley (2013). „MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.“ *Molecular Biology and Evolution* 30.4, pp. 772–780.

Lee, C.-M., L. Jiang, M. Whiteway, and S.-h. Shen (2004). „The serine/threonine protein phosphatase SIT4 modulates yeast-to-hypha morphogenesis and virulence in Candida albicans“. *Molecular Microbiology* 51, pp. 691–709.

Li, L., C. J. Stoeckert, and D. S. Roos (2003). „OrthoMCL: identification of ortholog groups for eukaryotic genomes.“ *Genome Research* 13.9, pp. 2178–2189.

Liebmann, B., T. W. Mühleisen, M. Müller, M. Hecht, G. Weidner, A. Braun, M. Brock, and A. A. Brakhage (2004). „Deletion of the Aspergillus fumigatus lysine biosynthesis gene lysF encoding homoaconitase leads to attenuated virulence in a low-dose mouse infection model of invasive aspergillosis“. *Archives of Microbiology* 181.5, pp. 378–383.

Link, T., C. Seibel, and R. T. Voegele (2014). „Early insights into the genome sequence of Uromyces fabae.“ *Frontiers in Plant Science* 5, p. 587.

Ma, T., J. Wang, G. Zhou, Z. Yue, Q. Hu, Y. Chen, B. Liu, Q. Qiu, Z. Wang, J. Zhang, et al. (2013). „Genomic insights into salt adaptation in a desert poplar.“ *Nature Communications* 4, p. 2797.

MacColl, A. D. (2011). „The ecological causes of evolution“. *Trends in Ecology & Evolution* 26.10, pp. 514–522.

Maier, W., B. D. Wingfield, M. Mennicken, and M. J. Wingfield (2007). „Polyphyly and two emerging lineages in the rust genera Puccinia and Uromyces.“ *Mycological Research* 111, pp. 176–185.

Marais, G, D Mouchiroud, and L Duret (2001). „Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes.“ *Proceedings of the National Academy of Sciences of the United States of America* 98.10, pp. 5688–5692.

Merwe, M. van der, L. Ericson, J. Walker, P. H. Thrall, and J. J. Burdon (2007). „Evolutionary relationships among species of Puccinia and Uromyces (Pucciniaceae, Uredinales) inferred from partial protein coding gene phylogenies.“ *Mycological Research* 111, pp. 163–175.

Miller, M., W. Pfeiffer, and T. Schwartz (2010). „Creating the CIPRES Science Gateway for inference of large phylogenetic trees“. In: *Proceedings of the Gateway Computing Environments Workshop (GCE)*, pp. 1–8.

Neafsey, D. E. and J. E. Galagan (2007). „Positive selection for unpreferred codon usage in eukaryotic genomes“. *BMC Evolutionary Biology* 7.1, p. 119.

Nemri, A., D. G. O. Saunders, C. Anderson, N. M. Upadhyaya, J. Win, G. J. Lawrence, D. a. Jones, S. Kamoun, J. G. Ellis, and P. N. Dodds (2014). „The genome sequence and effector complement of the flax rust pathogen Melampsora lini.“ *Frontiers in Plant Science* 5, p. 98.

Nery, M. F., D. J. González, and J. C. Opazo (2013). „How to Make a Dolphin: Molecular Signature of Positive Selection in Cetacean Genome.“ *PloS ONE* 8.6, e65491.

Pendleton, A. L., K. E. Smith, N. Feau, F. M. Martin, I. V. Grigoriev, R. Hamelin, C. D. Nelson, J. G. Burleigh, and J. M. Davis (2014). „Duplications and losses in gene families of rust pathogens highlight putative effectors“. *Frontiers in Plant Science* 5, p. 299.

Perpetua, N., Y Kubo, N Yasuda, Y Takano, and I Furuwasa (1996). „Cloning and characterization of a melanin biosynthetic THR1 reductase gene essential for appressorial penetration of Colletotrichum lagenarium“. *Molecular Plant-Microbe Interactions* 9, pp. 323–329.

Powell, S., K. Forslund, D. Szklarczyk, K. Trachana, A. Roth, J. Huerta-Cepas, T. Gabaldón, T. Rattei, C. Creevey, M. Kuhn, et al. (2014). „eggNOG v4.0: nested orthology inference across 3686 organisms“. *Nucleic Acids Research* 42.D1, pp. D231–D239.

Roux, J., E. Privman, S. Moretti, J. T. Daub, M. Robinson-Rechavi, and L. Keller (2014). „Patterns of Positive Selection in Seven Ant Genomes“. *Molecular Biology and Evolution* 31.7, pp. 1661–1685.

Ruiz-Herrera, J. (1994). „Polyamines, DNA methyltion, and fungal differentiation“. *Critical Reviews in Microbiology* 20.2, pp. 143–150.

Sela, I., H. Ashkenazy, K. Katoh, and T. Pupko (2015). „GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters“. *Nucleic Acids Research* 43.W1, W7–W14.

Shaffer, H. B., P. Minx, D. E. Warren, A. M. Shedlock, R. C. Thomson, N. Valenzuela, J. Abramyan, C. T. Amemiya, D. Badenhorst, K. K. Biggar, et al. (2013). „The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage.“ *Genome Biology* 14.3, R28.

Spanu, P. D. (2012). „The Genomics of Obligate (and Nonobligate) Biotrophs“. *Annual Review of Phytopathology* 50.1, pp. 91–109.

Stamatakis, A. (2014). „RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies“. *Bioinformatics* 30.9, pp. 1312–1313.

Staples, R. C. (2000). „Research on the Rust Fungi During the Twentieth Century“. *Annual Review of Phytopathology* 38.1, pp. 49–69.

Stergachis, A. B., E. Haugen, A. Shafer, W. Fu, B. Vernot, A. Reynolds, A. Raubitschek, S. Ziegler, E. M. Leproust, J. M. Akey, et al. (2013). „Protein Evolution“. *Science* 341.13, pp. 1367–1372.

Stukenbrock, E. and B. McDonald (2007). „Geographical variation and positive diversifying selection in the host-specific toxin SnToxA“. *Molecular Plant Pathology* 8.3, pp. 321–332.

Sun, C. B., A. Suresh, Y. Z. Deng, and N. I. Naqvi (2006). „A multidrug resistance transporter in Magnaporthe is required for host penetration and for survival during oxidative stress.“ *The Plant cell* 18.12, pp. 3686–3705.

Talavera, G. and J. Castresana (2007). „Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.“ *Systematic Biology* 56.4, pp. 564–577.

Talhinhas, P., H. G. Azinheira, B. Vieira, A. Loureiro, S. Tavares, D. Batista, E. Morin, A.-S. Petitot, O. S. Paulo, J. Poulain, et al. (2014). „Overview of the functional virulent genome of the coffee leaf rust pathogen Hemileia vastatrix with an emphasis on early stages of infection“. *Frontiers in Plant Science* 5, p. 88.

Tellier, A., S. Moreno-Gámez, and W. Stephan (2014). „Speed of adaptation and genomic footprints of host-parasite coevolution under arms race and trench warfare dynamics“. *Evolution* 68.8, pp. 2211–2224.

Tuller, T., Y. Y. Waldman, M. Kupiec, and E. Ruppin (2010). „Translation efficiency is determined by both codon bias and folding energy“. *Proceedings of the National Academy of Sciences* 107.8, pp. 3645–3650.

Valdés-Santiago, L., J. A. Cervantes-Chávez, R. Winkler, J. Ruiz-Herrera, and J. Ruiz-Herrera (2012a). „Phenotypic comparison of samdc and spe mutants reveals complex relationships of polyamine metabolism in Ustilago maydis“. *Microbiology* 158.3, pp. 674–684.

Valdés-Santiago, L., J. A. Cervantes-Chávez, C. G. León-Ramírez, and J. Ruiz-Herrera (2012b). „Polyamine metabolism in fungi with emphasis on phytopathogenic species.“ *Journal of Amino Acids* 2012, p. 837932.

Valle, M., H. Schabauer, C. Pacher, H. Stockinger, A. Stamatakis, M. Robinson-Rechavi, and N. Salamin (2014). „Optimization strategies for fast detection of positive selection on phylogenetic trees“. *Bioinformatics* 30.8, pp. 1129–1137.

Vamathevan, J. J., S. Hasan, R. D. Emes, H. Amrine-Madsen, D. Rajagopalan, S. D. Topp, V. Kumar, M. Word, M. D. Simmons, S. M. Foord, et al. (2008). „The role of positive selection in determining the molecular cause of species differences in disease.“ *BMC Evolutionary Biology* 8, p. 273.

Vitti, J. J., S. R. Grossman, and P. C. Sabeti (2013). „Detecting Natural Selection in Genomic Data“. *Annual Review of Genetics* 47.1, pp. 97–120.

Wang, X., S. D. Thomas, and J. Zhang (2004). „Relaxation of selective constraint and loss of function in the evolution of human bitter taste receptor genes.“ *Human Molecular Genetics* 13.21, pp. 2671–2678.

Wang, Y., G. Fan, Y. Liu, F. Sun, C. Shi, X. Liu, J. Peng, W. Chen, X. Huang, S. Cheng, et al. (2013). „The sacred lotus genome provides insights into the evolution of flowering plants.“ *The Plant journal* 76.4, pp. 557–567.

Wei, R.-X. and S. Ge (2011). „Evolutionary history and complementary selective relaxation of the duplicated PI genes in grasses.“ *Journal of Integrative Plant Biology* 53.8, pp. 682–693.

Wilson, R. A., J. M. Jenkinson, R. P. Gibson, J. A. Littlechild, Z.-Y. Wang, and N. J. Talbot (2007). „Tps1 regulates the pentose phosphate pathway, nitrogen metabolism and fungal virulence“. *The EMBO Journal* 26.15, pp. 3673–3685.

Win, J, W Morgan, J Bos, K. Krasileva, L. Cano, A Chaparro-Garcia, R Ammar, B. Staskawicz, and S Kamoun (2007). „Adaptive evolution has targeted the C-terminal domain of the RXLR effectors of plant pathogenic oomycetes.“ *the Plant Cell* 19.8, pp. 2349–2369.

Wu, M., S. Chatterji, and J. a. Eisen (2012). „Accounting For Alignment Uncertainty in Phylogenomics“. *PLoS ONE* 7.1, e30288.

Xia, X. (2013). „DAMBE5: A Comprehensive Software Package for Data Analysis in Molecular Biology and Evolution.“ *Molecular Biology and Evolution* 30.7, pp. 1720–1728.

Xia, X., Z. Xie, M. Salemi, L. Chen, and Y. Wang (2003). „An index of substitution saturation and its application“. *Molecular Phylogenetics and Evolution* 26.1, pp. 1–7.

Xiao, J.-H., Z. Yue, L.-Y. Jia, X.-H. Yang, L.-H. Niu, Z. Wang, P. Zhang, B.-F. Sun, S.-M. He, Z. Li, et al. (2013). „Obligate mutualism within a host drives the extreme specialization of a fig wasp genome.“ *Genome biology* 14.12, R141.

Yang, Z and J. Bielawski (2000). „Statistical methods for detecting molecular adaptation.“ *Trends in Ecology & Evolution* 15.12, pp. 496–503.

Yang, Z. (2007). „PAML 4: phylogenetic analysis by maximum likelihood.“ *Molecular Biology and Evolution* 24.8, pp. 1586–1591.

Yang, Z. and R. Nielsen (2002). „Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages.“ *Molecular Biology and Evolution* 19.6, pp. 908–917.

Zhang, J., R. Nielsen, and Z. Yang (2005). „Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level.“ *Molecular Biology and Evolution* 22.12, pp. 2472–2479.

Zhao, L., N. Zhang, P.-F. Ma, Q. Liu, D.-Z. Li, and Z.-H. Guo (2013). „Phylogenomic analyses of nuclear genes reveal the evolutionary relationships within the BEP clade and the evidence of positive selection in Poaceae.“ *PloS ONE* 8.5, e64642.

3

Population genomic footprints of host adaptation, introgression and recombination in Coffee Leaf Rust

Diogo N. Silva [1,2,3], Vítor Várzea [1], Octávio S. Paulo [2], Dora Batista [1,2]

[1] Centro de Investigação das Ferrugens do Cafeeiro, Instituto Superior de Agronomia, Universidade de Lisboa, Oeiras, Portugal.

[2] Computational Biology and Population Genomics group, cE3c – Centre for Centre for Ecology Evolution and Environmental Changes, Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal.

[3] Departamento de Biologia e CESAM - Centro de Estudos do Ambiente e do Mar, Universidade de Aveiro, Aveiro, Portugal.

3.1 Abstract

Coffee Leaf Rust, caused by *Hemileia vastatrix* (*Hv*), represents the biggest threat to coffee production worldwide and ranks amongst the most serious fungal diseases in history. Despite a recent series of outbreaks and emergence of hyper-virulent strains, the population evolutionary history and potential of this pathogen remains poorly understood. To address this issue, we used RADseq to generate 19 000 SNPs across a worldwide collection of 37 *Hv* samples. Contrarily to the longstanding

Published in Molecular Plant Pathology. DOI: 10.1111/mpp.12657

idea that *Hv* represents a genetically unstructured and cosmopolitan species, our results reveal the existence of a cryptic species complex with marked host tropism. Using phylogenetic and pathological data, we show that one of these lineages (*C3*) infects almost exclusively the most economically valuable coffee species (tetraploids that include *Coffea arabica* and inter-specific hybrids) while the other lineages (*C1-C2*) are severely maladapted to these hosts but successfully infect diploid coffee species. Population dynamic analyses suggest that the *C3* group may be a recent "domesticated" lineage that emerged via host-shift from diploid coffee hosts. We also found evidence of recombination occurring within this group, which could explain the high pace of pathotype emergence despite the low genetic variation. Moreover, genomic footprints of introgression between the *C3* and *C2* groups were discovered and raise the possibility that virulence factors may be quickly exchanged between groups with different pathogenic abilities. This work advances our understanding on the evolutionary strategies used by plant pathogens in agro-ecosystems with direct and far-reaching implications for disease control.

3.2 Introduction

In an era where modern ago-ecosystems create highly conducive environments for the appearance and dissemination of fungal pathogens, assessing the dynamics of these pathogens plays a crucial role in generating useful recommendations for disease management strategies (Stukenbrock and McDonald, 2008; Stukenbrock et al., 2011; Drenth and Guest, 2016; Gandon et al., 2016). However, if such strategies are to be effective and durable they must include eco-evolutionary principles that take into account the pathogen's evolutionary potential (Zhan et al., 2015; Grünwald et al., 2016). Pathogens with high evolutionary potential pose a greater risk of overcoming disease control strategies by usually having mixed reproductive systems, high gene flow, large effective population size and high mutation rates, features that may require special management practices (McDonald and Linde, 2002). In this regard, population genomics have provided an unprecedented amount of data to investigate the evolutionary processes that shape the structure of pathogen populations, pushing forward our understanding of pathogens' biology and usually

updating, or outright inverting, the prior assessment of the pathogen's ability to respond to control measures in agro-ecosystems (Milgroom et al., 2014; Talas and McDonald, 2015; Menardo et al., 2016).

Coffee Leaf Rust (CLR), caused by the obligate biotrophic fungus *Hemileia vastatrix* (*Hv*), is currently one of the biggest challenges to global coffee production and amongst the most serious crop diseases in history (Talhinhas et al., 2017). The disease causes premature defoliation on several species of the *Coffea* genus found at lower altitudes (<1000m) but only *C. canephora* and *C.arabica* are considered economically relevant at a global scale (McCook and Vandermeer, 2015). Of these two hosts, *C. arabica*, which is a recent tetraploid hybrid between the diploid *C. canephora* and *C. eugenioides* species, represents the most economically valuable species and also the most susceptible to *Hv* attacks. From the first epidemic report in Sri Lanka in 1869, *Hv* has spread to virtually every coffee-growing region in the world in about one and a half centuries (McCook and Vandermeer, 2015; Talhinhas et al., 2017). During this period several major outbreaks were registered in Asia and Africa and, more recently, a cluster of outbreaks has been occurring across the Americas since 2008 (Avelino et al., 2015; McCook and Vandermeer, 2015; Zambolim, 2016). Breeding for coffee rust resistance is considered to be the best long term solution to control CLR (McCook and Vandermeer, 2015), but the introduction of resistant varieties in the field has inevitably resulted in the loss of resistance due to adaptation of the pathogen. Coffee-rust interactions follow Flor´s gene-for-gene model (Flor, 1955) within a race-specific resistance system which imposes a coevolutionary "arms race". The continuous exertion of selective pressure on the pathogen by the resistant varieties has led to the emergence of more than 50 races or pathotypes, which is remarkable for a supposedly asexual pathogen (Talhinhas et al., 2017). As of now, hyper virulent CLR isolates able to infect coffee genotypes previously resistant to all known rust pathotypes have already been identified in India (Prakash et al., 2014).

The seriousness of CLR epidemics has triggered emergency actions across coffee producing nations and investigation of the pathogen's biology is gaining a considerable momentum with attempts to gather genomic (Cristancho et al., 2014) and

transcriptomic data (Talhinhas et al., 2014). However, considerably less explored remains the evolutionary history of *Hv* populations, particularly at a global scale. Using RAPDs and AFLP markers, population genetic studies of *Hv* populations have predominantly focused on restricted geographical areas, such as Brazil (Nunes et al., 2009; Maia et al., 2013; Cabral et al., 2016) and Colombia (Rozo et al., 2012), with the exception of Gouveia et al. (2005) that included isolates from Asia, Africa and America. These studies have consistently found no evidence of population structure with respect to pathotype, host or geographical origin but they report mixed results regarding the level of genetic variability. The sexuality of *Hv* is also a debatable subject in the literature, but *Hv* is generally considered to be an asexual pathogen due to the fact that the sexual phase of its life-style was not identified so far and that the asexual urediniospores are the only known functional propagules (Silva et al., 2006; Talhinhas et al., 2017). However, meiosis was recently discovered within the urediniospores in a supposedly hidden sexual reproductive cycle (Carvalho et al., 2011). Whether this means that recombination effectively occurs in natural populations of *Hv*, remains an open question. Some studies were unable to detect recombination and support the asexual status of *Hv* (Gouveia et al., 2005; Rozo et al., 2012), while others have found evidence of recombination is some specific regions (Maia et al., 2013; Cabral et al., 2016).

In this study we used RAD sequencing to generate thousands of molecular markers for *Hv* with the goal of investigating its genetic structure and population dynamics from a worldwide sample collection that includes a broad range of pathotypes and isolates infecting several coffee species. Specifically, our aims were to: (i) produce a large SNP data set of high quality by controlling sequencing error using individual sample replicates; (ii) investigate how genetic variation within *Hv* populations is distributed according to host, race pathotype and geographical origin; (iii) test for the presence of recombination within *Hv*.

3.3 Materials and methods

3.3.1 Fungal material, sample preparation and RAD sequencing

Twenty nine isolates of *H. vastatrix* (*Hv*) from 11 geographical locations, collected from different diploid and tetraploid coffee hosts and comprising 18 unique pathotypes were retrieved from a spore collection maintained at CIFC (Table A.2.1). Pathotypes were determined based on inoculation assays on a set of coffee differentials bearing different resistance gene combinations under standard testing conditions (D'Oliveira, 1954). Coffee differentials include tetraploid coffee plants with resistance genotypes containing the nine major dominant genes (S_H1 to S_H9), either individually or in combination, identified through classical genetics according to Flor's gene-for-gene model (Bettencourt and Rodrigues, 1988), and also six diploid coffee hosts (*C. racemosa*, *C. excelsea*, two *C. canephora* and two *C. congensis*). According to the virulence profile on the differential plants, isolates are classified into pathotypes (races) comprising virulence genes as inferred by Flor's theory, ranging from v_1 to v_9 in isolates derived from *C. arabica* and tetraploid interspecific hybrids, whereas those of the races that attack diploid coffee species are not known ($v_?$). Individual sample replicates were added for 9 isolates, chosen in order to encompass the largest possible range of geographical locations, pathotypes and hosts. The total sampling contained 38 isolates that were processed independently. DNA was extracted using a CTAB-based protocol modified from Kolmer et al. (1995) and genomic DNA concentration and quality was checked by visual inspection on an agarose gel and with a ND-1000 Nanodrop spectrophotometer. Three micrograms of high quality genomic DNA per individual were sent to Floragenex Inc. (Oregon USA) for RAD library preparation and sequencing as previously described (Etter et al., 2011). Libraries with sample-specific barcode sequences were produced from DNA digested with *PstI* enzyme and single end (1x100bp) sequencing was performed in an Illumina HiSeq 2000 machine.

3.3.2 RADseq assembly strategy and SNP calling

Sequence reads were de-multiplexed and quality filtered with the PROCESS_RADTAGS script (Catchen et al., 2013). Reads with uncalled bases or distance to barcodes higher than 1 were removed. Base calls with a Phred score under 20 were converted to Ns and reads containing more than 4 Ns were discarded. RAD tags were *de novo* assembled and genotyped using PyRAD v3.0.63 (Eaton, 2014) due to its ability to handle gaps when clustering sequence reads. In order to optimize the assembly process, we used a similar strategy as described in (Mastretta-Yanes et al., 2015), in which individual sample replicates were used to assess error rates across a range of values for three major assembly parameters. Four error rates were considered for each assembly: (i) total locus error, (ii) partial locus error, (iii) haplotype error and (iv) SNP error. A complete description of the specified error rates is provided in section 3.6.1 (page 108). Several values were tested for each of the following three assembly parameters separately, while fixing the remaining parameters: (i) minimum read depth (5-15); (ii) clustering threshold (0.80-097) and (iii) maximum shared heterozygosity (2-10). An overview of the parameter values for each assembly is provided in Table A.2.2. Error rates and total number of SNPs were evaluated for each assembly, after the removal of SNPs with less than 50% of taxa representation, using the custom COMPARE_PAIRS.PY script. The assembly that yielded the lowest error rates while maximizing the total number of SNPs was then processed for further analyses. SNPs with a mismatch in at least one replicate were removed from the VCF file. Replicates were also removed, retaining whichever isolate from the pair that contained more data. Handling and exploration of alignment data matrices was performed using TriFusion v0.4.12 software (https://github.com/OdiogoSilva/TriFusion).

3.3.3 Phylogenetic analyses

To infer the phylogenetic relationships among our samples we applied a supermatrix approach that included loci with SNPs represented in more than 50% of the isolates and a minor allele frequency above 5% into a single concatenated align-

ment. Concatenation and conversion of the alignment matrices to the appropriate formats was performed with TriFusion. Maximum likelihood (ML) reconstruction was performed using RAxML v8.2.6 (Stamatakis, 2014) with the `GTRCAT` model of sequence evolution and with bootstrap support estimated from 1 000 replicates. The same data matrix was used for phylogeny estimation using a Bayesian framework as implemented in MrBayes v3.2.6 (Ronquist et al., 2012) with the `GTR` + Γ model of sequence evolution. Posterior probabilities were generated from 1 x 10^7 generations, sampling at every 1 000^{th} iteration, and the analysis was run three times with one cold and three incrementally heated Metropolis-coupled Monte Carlo Markov chains, starting from random trees. The achievement of the stationary phase and mixing was checked for all parameters using Tracer v1.4, and 1 x 10^6 generations were discarded as burn-in. Trees from different runs were combined using LogCombiner and summarized in a majority rule 50% consensus tree. Both RAxML and MrBayes runs were performed in the Cipres Science Gateway clusters (Miller et al., 2010).

3.3.4 Evaluation of *H. vastatrix* genetic structure

To detect potential admixture and population structure in our *Hv* sampling, the software fastSTRUCTURE v1.0 (Raj et al., 2014) wrapped in Structure_threader v0.4.3 (Pina-Martins et al., 2017) was used with K values ranging between 1-8. The optimal K value was found using the chooseK.py script bundled in Structure_threader. Principal Component Analysis (PCA) was run using the SNPRelate v1.8.0 R package (Zheng et al., 2012), after filtering non-biallelic loci, using the `snpgdsPCA` function. The degree of population differentiation was also assessed by calculating the overall and distribution of SNP F_{ST} values for each population pair using VCFTOOLS v0.1.14 (Danecek et al., 2011).

3.3.5 Introgression assessment

Early on this work, multiple genetically differentiated groups were detected within our sampling with some potential evidence of introgression found for specific isolates. To better assess this potential signal of introgression at the individual level, and

to exclude the possibility of incomplete lineage sorting, we devised a simple SNP scanning strategy between the two groups of isolates. First, we filtered the SNP data set so that only SNPs with an F_{ST} value above 0.8 were retained. These were named “diagnostic” SNPs, since they were able to almost completely differentiate both groups, while allowing the sharing of alleles for very few isolates (presumably the isolates with introgressed loci). Then, for each isolate in one of the groups, we scanned each “diagnostic” SNP and kept score of how many alleles were shared with the genotype of the other group and whether the shared alleles were homozygous or heterozygous. Our expectation for this assessment was that isolates with a signal of introgression would contain much higher numbers of shared alleles than non-introgressed taxa.

3.3.6 Recombination and linkage disequilibrium

The linkage disequilibrium (LD) metrics D' and r^2 were used to estimate the pair-wise LD between all SNPs that passed the MAF filter. Both metrics were calculated using the `LD` function of the **GENETICS** v1.3.8.1 R package, which is able to estimate the proportion of heterozygous genotypes using ML. In addition, a χ^2 statistic test was calculated for each pair, providing a p-value to test for marker independence. In addition, the fact that fungal populations may not be strictly clonal or sexual but are often found in an intermediate position was taken into account by estimating the standardized index of association ($\bar{r}_d$) (Agapow and Burt, 2001), as implemented in the **POPPR** v2.3.0 R package (Kamvar et al., 2014). The index of association tests to what extent individuals that are the same at one locus are more likely than random to be the same at other loci, and is thus a measure of linkage disequilibrium based on the variance of pairwise distance between individuals. The standardized form of the Index of Association, $\bar{r}_d$, provides a more unbiased test that is independent of the loci sample size. In order to obtain an expected null distribution and a p-value, 999 permutations were performed on the data.

3.3.7 Population dynamics of *H. vastatrix* infecting tetraploid hosts

The demographic history of *Hv* populations was reconstructed using the Extended Bayesian Skyline model (Heled and Drummond, 2008) implemented in **BEAST** v1.8.3 (Drummond et al., 2012). The data set for this analysis consisted only of RAD loci that were present in all analysed isolates (no missing data) and was assembled using **TriFusion**. A single molecular clock partition was established using the uncorrelated lognormal prior. Since there are no calibration points available within our data set, the arbitrary starting value of 1 was chosen for the `ucld.mean` parameter of the clock model. The analysis was run twice, with default priors, for 1 x 10^8 generations, sampling at every 10 000^{th} generation after an initial burn-in of 10%. The performance of the MCMC procedure, namely the ESS values and mixing for each parameter, was assessed in **Tracer** v1.6 (Rambaut et al., 2014).

3.4 Results

3.4.1 RAD-Seq data assembly and quality control

Illumina RAD sequencing of 38 *H. vastatrix* (*Hv*) isolates from 11 geographical locations and different coffee hosts, and comprising 18 unique pathotypes generated an average of 4.48x10^6 reads per sample, amounting to 4.09x10^8 base pairs per sample (Table A.2.3). From the total sampling, 9 isolates consisted of technical replicates used to test assembly parameters. Eleven *de novo* assemblies were performed and the results are summarized in Table A.2.4. The best assembly strategy resulted in a total locus error of 40.40%, partial locus error of 12.52%, allele error of 4.64% and SNP error of 3.92% and yielded a final matrix of 19 505 SNPs across 14 556 loci and 29 isolates. The additional filter of excluding SNPs with a minor allele frequency (MAF) lower than 5% reduced the data set to 8 389 SNPs (6 783 phylogenetically informative) across 6 783 loci and 29 isolates.

3.4.2 Phylogenetic analysis

Phylogenetic reconstruction of the 29 *Hv* isolates with a concatenated data set of 6 783 variable loci using ML and Bayesian methods yielded similar topologies with congruent branch support values (Figure 3.1). The resulting best tree revealed the presence of three divergent and well supported groups (*C1*-*C3*), which appear to be highly structured according to the host species. The eight isolates from groups *C1* and *C2* were sampled from diploid coffee species (*C. canephora*, *C. racemosa*, *C. liberica* and *C. excelsa*), whereas the remaining 21 isolates from group *C3* were collected from *C. arabica* and inter-specific hybrids, all of which are tetraploid. No other clear sub-structuring was observed concerning the geographical location or pathotype of the isolates. The only exception is a shallow but well-supported five isolate group within the *C3* group that includes the most basic pathotypes (v_5 and $v_{1,5}$). In the basal position of the *C3* group, there are three isolates of different geographic locations and pathotypes in a ladder-like pattern (*999*, *2377* and *3624*).

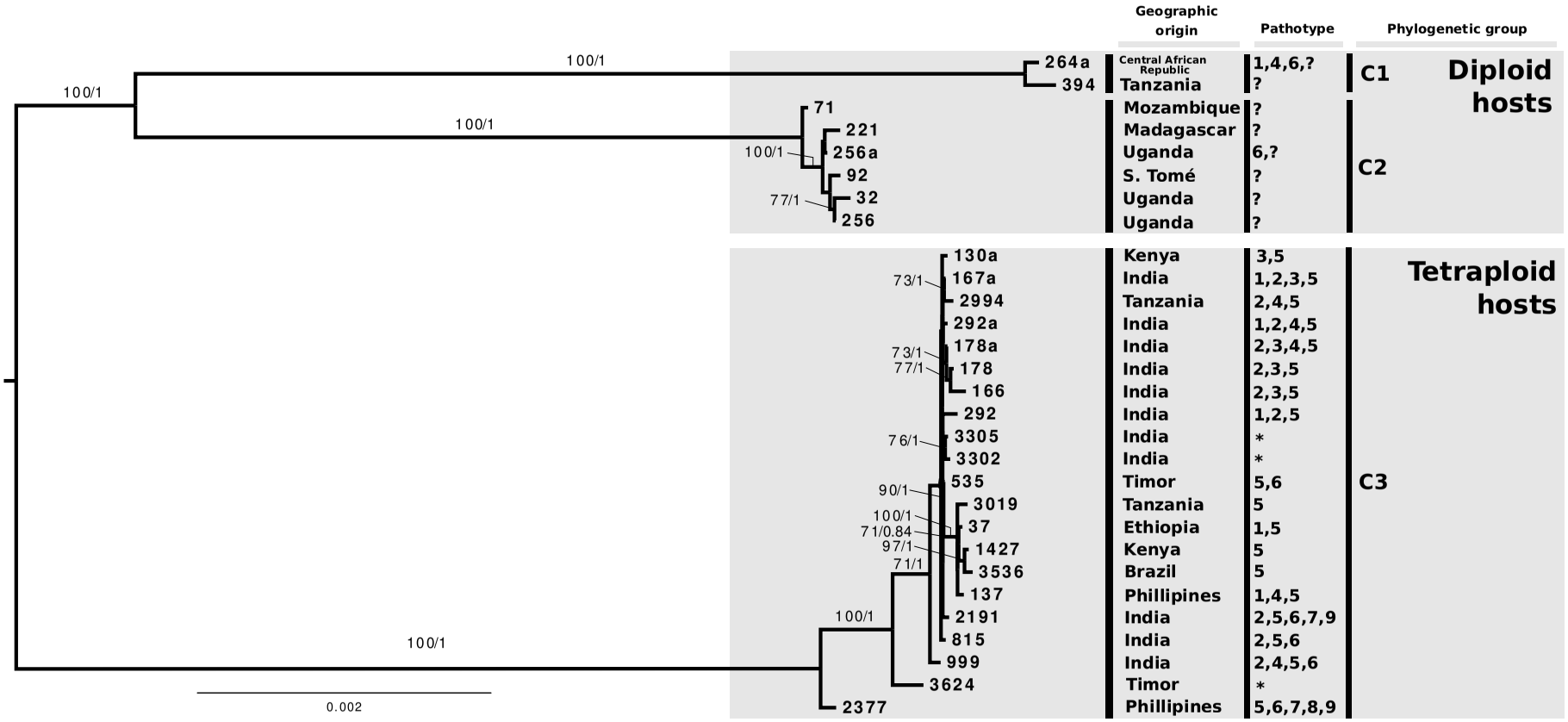

Figure 3.1. Phylogenetic relationships among 29 isolates of *H. vastatrix*. Support values are provided above branches with bootstrap values above 70 and posterior probability above 0.8. For each isolate, information about its geographic origin, pathotype and phylogenetic group is provided.

3.4.3 Population structuring of *H. vastatrix*

FastSTRUCTURE analyses revealed that the most likely number of clusters (K) in our data set was three, which corresponds to the same three clusters identified in the phylogenetic analyses (Figure 3.2). For the majority of the isolates, the membership coefficient was at the maximum value for the corresponding cluster, indicating a substantial degree of population differentiation.

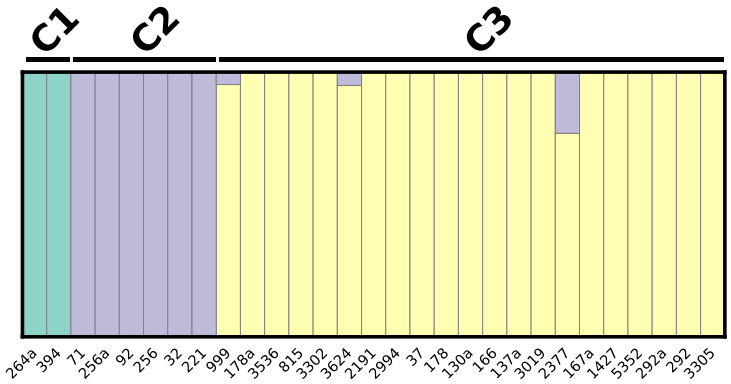

Figure 3.2. Structure plot of the 29 *H. vastatrix*'s isolates with K=3. Vertical bars represent an isolate and the colour proportion for each bar represents the posterior probability of assignment to one of the three clusters. The three groups identified in the phylogenetic tree are outlined above the plot.

However, an admixture signal was detected for three isolates of the *C3* group, which correspond to the isolates found at the base of the *C3* phylogenetic group (*999*, *2377* and *3624*). This signal suggests that, for those isolates, a proportion of loci was more closely associated with the *C2* group (infecting diploid hosts) than with its own group. The PCA analysis also revealed the same pattern of three clusters as the previous analyses, with the first and second principal components explaining 35.37% and 20.34% of the variance, respectively (Figure 3.3). It is noteworthy that while most of the isolates from the *C3* group are very closely clustered, five isolates (*292*, *999*, *166*, *3624* and *2377*) appear to stand apart from the group. Again, three of these isolates correspond to the same isolates with admixture signal in the **fastSTRUCTURE** analysis and at the base of the *C3* group phylogenetic tree. Estimates of F_{ST} values for each population pair further supported a near complete genetic differentiation between each of the three groups, with weighted F_{ST} values ranging from 0.92 (*C1* x *C2*) to 0.95 (*C1* x *C3* and *C2* x *C3*) (Figure A.1).

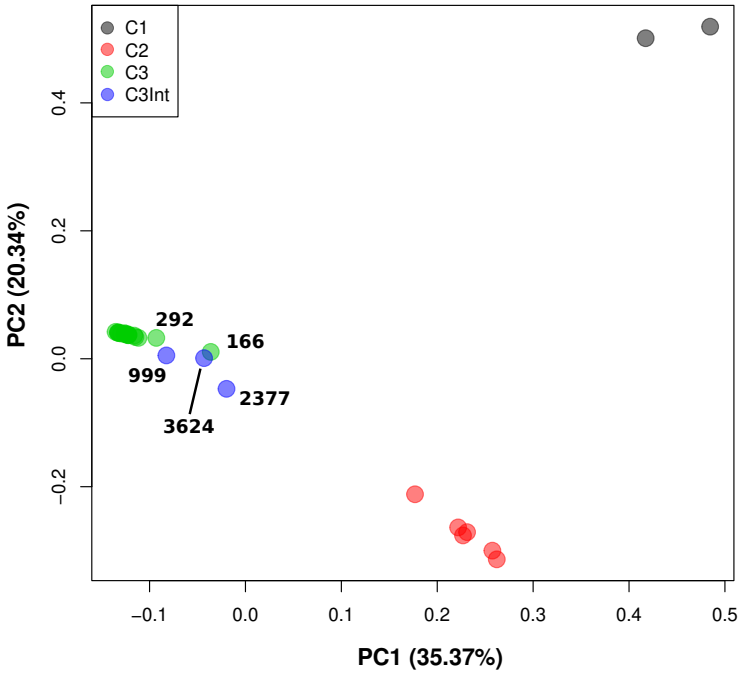

Figure 3.3. Principal component analysis of genomic diversity for 29 isolates of *H. vastatrix*. Isolates are colour coded according to their assignment to the three phylogenetic groups. The three isolates of the *C3* group that revealed a signal of allele sharing with the *C2* group were further differentiated as a fourth *C3Int* group.

3.4.4 Genetic diversity

For the estimation of genetic diversity indexes, we focused only on the *C2* and *C3* groups due to the low number of samples of the *C1* group (2 isolates). From a total of 19 505 SNPs from the full data set, only 2 831 were segregating within the *C2* group and 7 503 within the *C3* group. When we applied a MAF filter that allowed only SNPs with more than one allele for each group (5% for the *C3* group and 18% for the *C2* group), the number of segregating SNPs was further reduced to 551 in *C2* and 1 563 in *C3*. This result was further corroborated by the allele frequency spectrum of each group, which revealed a distribution markedly skewed to lower frequencies of derived alleles (Figure A.2), and by the high proportion of singletons in *C2* (91.81%) and C3 (90.03%) groups. Considering the inbreeding coefficient, F_{IT}, for each isolate in the *C2* and *C3* groups, we found that the majority of the isolates had positive values ranging from 0.18 to 0.98, which indicates a slight to moderate excess of homozygotic SNPs compared to the theoretical expectation (Figure 3.4). Notable exceptions were the three isolates previously found at the base of the *C3* phylogenetic group and showing the allele sharing signal, which

consistently revealed negative F_{IT} values, indicating an excess of heterozygotic SNPs.

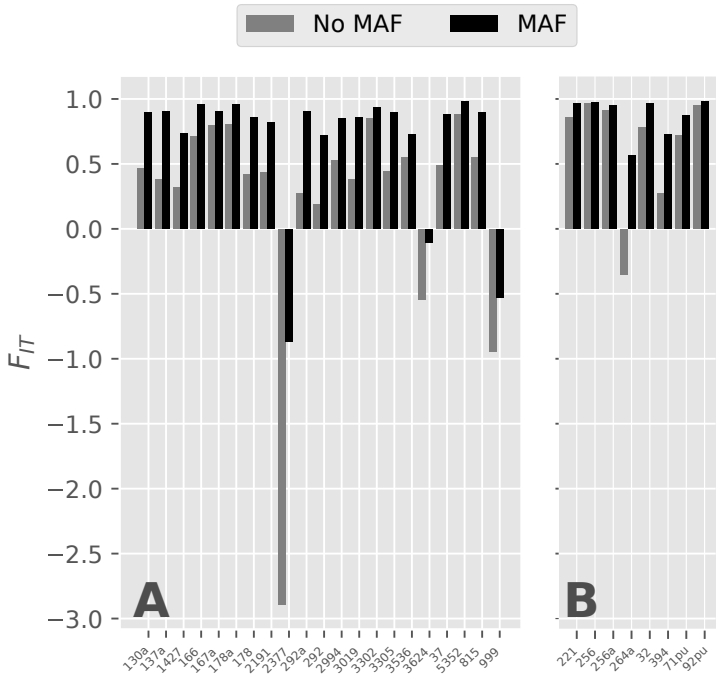

Figure 3.4. Bar plots of the inbreeding coefficient (F_{IT}) for each isolates of *H. vastatrix* from the *C3* (A) and *C2* (B) groups. For each isolate, F_{IT} values were calculated for data sets with (MAF) and without (No MAF) a minor allele frequency filtering.

3.4.5 Investigating introgression between *H. vastatrix* groups

The allele sharing between the *C3* and *C2* groups was investigated for each isolate by scanning 4 494 diagnostic SNPs. The results, summarized in Figure 3.5, clearly reveal that most isolates of the C3 group have very few shared alleles with isolates of the C2 group, except for isolates 3624, 999, and 2377 (Figure 3.5).

Indeed, from a total of 1 938 SNPs that displayed at least one shared allele, 1 619 (83.53%) were found exclusively in the three admixed isolates. Individually, these isolates revealed 5.83% (3624), 12.29% (999) and 30.65% (2377) of SNPs with shared alleles with the *C2* group with the vast majority of these SNPs being heterozygous. In contrast, these SNPs were mostly homozygous for the remainder isolates in their respective groups. We also verified if there was a consistent signal of

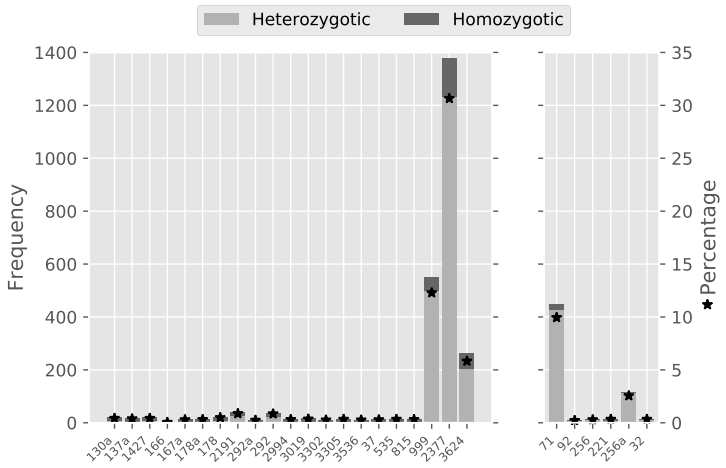

Figure 3.5. Summary of the diagnostic SNP scanning for shared alleles between isolates of the *C2* and *C3* phylogenetic groups. The stacked bar plot represents the frequency, while the star point plot represents the percentage, of alleles that isolates shared with the other group.

allele sharing across loci that contain two or more SNPs, which would be expected due to the linkage generated from their proximity. Indeed, between 99.02% and 99.58% of the SNPs in these loci agreed on the allele sharing signal. We also assessed the overlap of the SNPs with shared alleles between these three isolates and found that only a small proportion were shared among the three or in pair combinations (Figure 3.6).

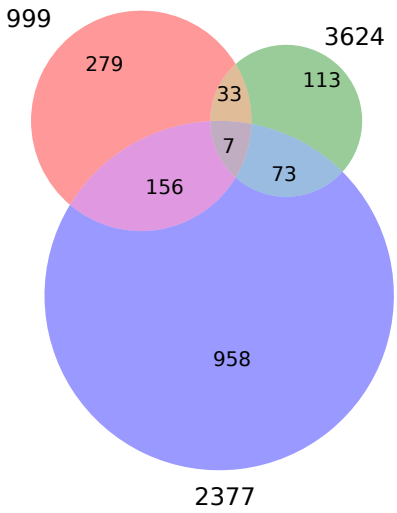

Figure 3.6. Venn diagram with the overlap of the SNPs with shared alleles among the three *H. vastatrix* isolates showing admixture signal.

For instance, only seven SNPs with shared alleles were found in all three admixed isolates and, among the pair-wise combinations, the overlap was never higher than

33%. The majority of the SNPs with shared alleles were exclusive to each of the three isolates (2377: 80.45%; 999: 58.80%; 3624: 50.22%). These analyses were also repeated with different F_{ST} values used to generate the "signature" SNPs (0.5 and 0.9) and the results were qualitatively equal and quantitatively similar. Finally, we assessed if the introgression events occurred in both directions by performing these analyses considering the *C2* group as the recipient. Even though previous analyses did not hint at the presence of allele sharing in that direction, our SNP scan detected two *C2* isolates with 9.96% (71) and 2.56% (256a) SNPs with shared alleles with the *C3* group, the majority being heterozygous (Figure 3.5). The remaining *C2* isolates displayed only vestigial amounts of shared SNPs.

3.4.6 Linkage disequilibrium and recombination

To estimate Linkage Disequilibrium (LD) metrics and investigate the presence of recombination, we focused on the larger and epidemiological more relevant *C3* group. The mean value of D' across SNP pairs was high (0.86±0.27) and mean r^2 was low (0.13±0.25) (Table A.2.5). Out of the total pair-wise SNP comparisons, only 17.8% SNP pairs could reject the null hypothesis of no association between genotypes, that is, of being in linkage disequilibrium. Given the presence of isolates with a putative signal of introgression from the *C2* group and the existence of a *C3* sub-group containing isolates with basal pathotypes, we created one additional data set where the three introgressed isolates were removed (`NoInt`; n=18) and another where the isolates from the *C3* sub-group were also removed (`NoInt_NoV5`; n=13). These data sets were meant to remove potential biases that introgression and mild population structuring could introduce in genotype association, since these phenomena are known to inflate estimates of linkage disequilibrium even in the presence of recombination (Agapow and Burt, 2001). The overall values of D' and r^2 remained similar for the new data sets, but the percentage of SNP pairs rejecting the null hypothesis decreased to 7.73% in the `NoInt` data set ($D' = 0.87 \pm 0.29$; $r^2 = 0.08 \pm 0.21$) and 10.98% in the `NoInt_NoV5` data set ($D' = 0.89 \pm 0.28$; $r^2 = 0.26 \pm 0.26$) (Table A.2.5). We then estimated the standardized form of the index of association ($\bar{r}_d$) for the same data sets in order to assess whether the patterns of genetic variation in the *C3* group were consistent with clonal or sexual reproduction.

When using the complete *C3* group (including putatively introgressed isolates and basal pathotypes) $\bar{r}_d$ was low (0.05) but significantly higher than the expected null distribution of no linkage among markers ($p = 0.001$) (Figure A.3). The same result was obtained for the `NoInt` data set ($\bar{r}_d = 0.02$; $p = 0.001$) (Figure A.3), but when both introgressed taxa and isolates from the *C3* sub-group were removed, the index of association could not reject the null hypothesis of sexual reproduction ($\bar{r}_d = 0.01$; $p = 1$) (Figure 3.7).

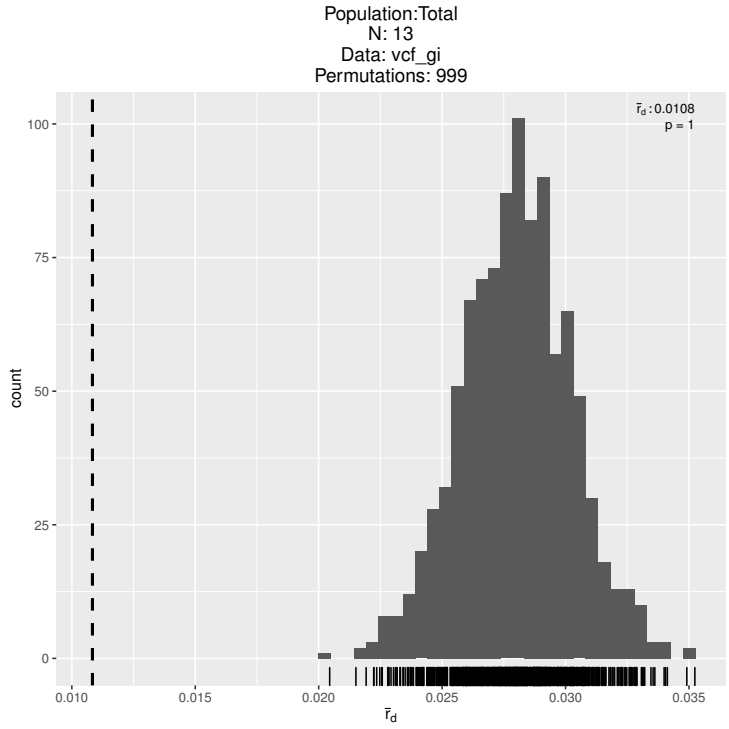

Figure 3.7. Results from the Index of Association (IA) analysis, using the standardized form (r_d), for the isolates of the *C3* group after removing putative introgressed isolates and an incipient but well supported sub-group. The histogram depicts the distribution of r_d values expected from unlinked loci. The vertical dashed line represents the observed rd value for the data set.

3.4.7 Population dynamics of *H. vastatrix* in tetraploid hosts

The demographic history of *Hv* isolates from the *C3* group was reconstructed using an extended Bayesian Skyline analysis (Figure 3.8). Since there was no calibration available for our data set, time estimates should be interpreted in relative terms. The historical demographic reconstruction revealed that the *C3* group suffered a severe bottleneck during and shortly after the divergence from the remaining groups infecting diploid hosts (*C1-C2*). After this bottleneck event, the effective

population size seems to have remained low for most of the time until the onset of the diversification of the *C3* group, at which point an increase in population size to values above pre-bottleneck is inferred (Figure A.5). We should note that the 95% High Posterior Density intervals for these estimates can be quite large, particularly for the size of the expansion in recent times. However, the pattern of ancient bottleneck with a recent population expansion still holds, even when different model specifications were experimented during the Skyline analyses.

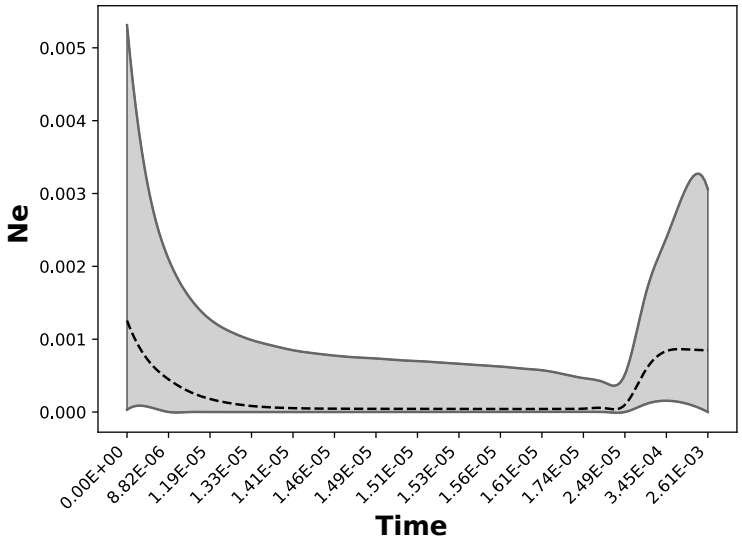

Figure 3.8. Extended Bayesian skyline plot depicting the population dynamics of the *C3* group of *H. vastatrix* through time. The x-axis is in relative unites of time, and the y-axis corresponds to the effective population size. The dashed black line represents the median estimate of the effective population size, while the solid grey lines delimit the 95% high posterior density.

3.5 Discussion

3.5.1 Revealing *H. vastatrix* as a potential cryptic species complex with host specialization

One of the first and most striking findings of this work was the discovery of three well diverged evolutionary lineages within our sampling of a supposedly single species, which is in disagreement with previous studies (Rozo et al., 2012; Maia et al., 2013; Cabral et al., 2016). Until now, disease management practices were heavily influenced by the idea that *Hv* represents a large unstructured population

that is able to freely switch between coffee species (McCook and Vandermeer, 2015; Zambolim, 2016). By contrast, our results show a clear phylogenetic segregation between isolates infecting *C. arabica* and inter-specific tetraploid hybrids (which will henceforth be collectively referred to as tetraploids), and isolates infecting *C. canephora* and other diploid coffee species with no commercial relevance. This structuring pattern is maintained even when an extended sampling of *Hv* ($n = 119$) was assessed in preliminary analyses of an ongoing project. The pathogenicity tests routinely performed at CIFC also support this segregation, at least when considering the commercial coffee plants. Some coffee species of little or no commercial importance, such as *C. racemosa* and *C. liberica*, are universally susceptible to all *Hv* isolates, but their influence in the pathogen's dynamics is low due to their limited geographic distribution. However, this is clearly not the case for tetraploids and *C. canephora*. Isolates from the *C3* group are consistently unable to infect several *C. canephora* differential genotypes, whereas isolates from the *C1* and *C2* groups are either unable to infect tetraploids, or trigger only mild symptoms with limited spore production in a few varieties. The immediate deduction from this is that adaptation to either tetraploids or *C. canephora* entails an adaptive trade-off for *Hv*: the ability to successfully infect one host implies poor or no fitness in the other. This marked host tropism is quite intriguing when we consider that, at the cytological level, the infection process of compatible interactions is similar across all *Hv* isolates, regardless of the host (Silva et al., 2008). Indeed, without the knowledge of *Hv*'s phylogenetic structure there would be little reason to consider the output of cross-inoculation tests anything other than the presence of different *Hv* races. Coincidentally, this would also explain the longstanding perception that varieties of *C. canephora* are inherently more resistant to CLR (McCook and Vandermeer, 2015; Talhinhas et al., 2017). What seems to be the case, however, is that *C. canephora* is highly resistant to isolates adapted to tetraploids, which happen to represent the most widespread and epidemiological relevant group of *Hv*.

Given the implications that the discovery of multiple divergent and pathologically different lineages within *Hv* would have in understanding its evolutionary potential, we further investigated the genetic differentiation of these groups. Results from clustering analyses were unanimous in the confirmation of the three groups identified

in the phylogenetic analysis, and further highlighted their high degree of genetic differentiation. With average weighted F_{ST} values above 0.90 in all pair-wise group comparisons, they seem to be almost completely isolated from each other at the genetic level. Notwithstanding, not only is the infection process identical across all isolates, but they are also morphologically similar (Silva et al., 2006). This raises the question of whether these groups should be considered cryptic species. Cryptic species complexes are widespread in the fungal kingdom (Pringle et al., 2005; Silva et al., 2012b; Stukenbrock, 2013), including rust fungi (Bennett et al., 2011; Zhao et al., 2015). In fact, agro-ecosystems have favoured the emergence of new and specialized closely related species by providing a new ecological niche (Silva et al., 2012a; Stukenbrock and Bataillon, 2012), and the correct identification of species boundaries can be difficult even with the aid of molecular data. However, in the case of *Hv*, there are implications that go beyond the eventual update of the taxonomic status for these groups. From an epidemiological standpoint, it is crucial to understand to what extent these groups are genetically isolated.

3.5.2 Investigating the presence of introgression between groups of *H. vastatrix* infecting diploid and tetraploid hosts

Despite the high genetic differentiation, evidence of some allele sharing between the *C2* and *C3* groups was revealed early in our analyses. The challenge here was to assess whether this signal was the result of incomplete lineage sorting or introgression, since both can produce similar patterns of allele sharing (Twyford and Ennos, 2012). Similar to hybridization and introgression, incomplete lineage sorting can cause an excess of shared derived alleles but this excess arises from purely non-contemporary demographic phenomena, such as population fragmentation or non-random mating in the ancient population (Eriksson and Manica, 2012). Nevertheless, these processes leave distinct signatures in the patterns of the genetic diversity of populations that can be unravelled by genome-wide data (Staubach et al., 2012; Twyford and Ennos, 2012). For instance, considering the *C3* group, our fine-scale assessment of diagnostic SNPs revealed that as much as 84% of the SNPs with

alleles shared with the *C2* group were found exclusively in three isolates (*3624*, *999* and *2377*). This markedly skewed distribution towards only a few isolates is coupled with the observation that virtually all SNPs in physical linkage have the same allele sharing signal. Contrarily to the expectations of incomplete lineage sorting, where allele sharing is expected to be unlinked and reasonably distributed across all isolates, this is the expected result from introgression events. In this scenario, entire chromosomal segments of one species are inserted into the genetic background of the other species, creating blocks of linked SNPs found only on the introgressed offspring (Rheindt et al., 2014). The patterns of heterozygosity in SNPs with shared alleles also lend support to introgression over incomplete lineage sorting. The vast majority of the SNPs with shared alleles were homozygous for all non-admixed isolates of the *C2* and *C3* groups but heterozygous for the three admixed isolates. It is hard to conceive a scenario where purely demographic phenomena would lead to the fixation of alternate variants in two populations, with the consistent exception of a few isolates across hundreds of variable sites. On the other hand, hybridization and subsequent introgression are expected to produce this pattern of genetic variation (Harrison and Larson, 2014; Todesco et al., 2016). First (*F1*) or early generation hybrids will be heterozygous for nearly all sites segregating between the two parental species. Backcrossing and introgression of the hybrids into one of the parental species would result in heterozygous variants as observed in the three admixed isolates. The introgression signal in the admixed isolates could also explain their basal position in the respective group, even thought their pathotypes are highly derivative. It is widely known that hybridization and introgression can interfere with phylogenetic reconstructions, as introgressed individuals will appear at intermediate positions between the parental groups, depending on the extension of the introgression (Eaton and Ree, 2013).

Overall, even though we did not detect any *C2*x*C3 F1*-hybrids in our sampling, introgression between the *C2* and *C3* groups appears to be the most likely explanation for the shared polymorphism pattern in our data. Moreover, the presence of alleles in the *C2* group that are shared with the *C3* group suggests that introgression may be bi-directional. This has major implications for the epidemiology of *Hv* as it creates the possibility of genetic exchange between two divergent groups adapted to

different sets of coffee hosts. In principle, this could quickly generate novel genetic recombinants and promote rapid adaptation to new resistant coffee hosts. Hybridization and introgression is being increasingly regarded as an important venue for the rapid emergence of novel pathotypes or even species (Stukenbrock and McDonald, 2008; Menardo et al., 2016; Stukenbrock, 2016). In the case of *Hv* this would certainly represent an opportunity for the pathogen to respond to the introgression of resistance genes from diploid coffee species into tetraploids. However, whether these introgression events are actually adaptive or not remains untested. In our sampling we could only detected a clear signal of introgression in three isolates of the *C3* group, but it is entirely possible that more events have already occurred. Our results are not consistent with this being a single and geographically restricted event, but most likely the result of multiple independent events. The introgressed isolates were found in geographically distant locations and present a very small overlap in the introgressed SNPs. Principal component analysis of the *C3* group alone also reveal that the introgressed isolates do not cluster together (Figure A.4). If we combine this information with the occurrence of recombination within the *C3* group (see below) the dissemination of only a few adaptive key alleles that are harder to detect becomes a likely possibility – and one that may be more common than previously thought (Anderson et al., 2009; Rheindt et al., 2014).

3.5.3 The emergence and evolutionary history of the C3 group

With the clustering of isolates infecting the most economically relevant coffee hosts in a single group, our interest was then shifted to how the *C3* group emerged and subsequently evolved. Even with the application of thousands of SNPs, no significant genetic structuring was found according to geographical origin or pathotype within this group, besides the incipient separation of isolates with basal pathotypes. To explain this apparent panmixia, in addition to the previously identified factors of worldwide expansion of coffee trade and high vagility of *Hv*'s spores (McCook and Vandermeer, 2015; Talhinhas et al., 2017), this work suggests two additional factors that are tightly linked to the evolutionary potential of the *C3* group: a recent origin

and reticulate evolution, that is, evolution shaped by hybridization/recombination between diverging lineages.

Since the genetic structure of the *C1-C3* groups was unknown so far, it was never considered that the emergence of *Hv* in tetraploid hosts could have been the result of a recent introduction and adaptation event. Notwithstanding, this would be consistent with the also recent origin of *C. arabica* from a hybridization event between *C. eugenioides* and *C. canephora* (Cenci et al., 2012). The reconstruction of the *C3* group's population dynamics lends support to this hypothesis by revealing a pattern of an early bottleneck followed by a relatively long stable period that culminated in a very recent population size explosion. The timing of the inferred bottleneck coincides with the divergence time between the *C3* and *C1-C2* groups and the population expansion closely matches the onset of the *C3* group diversification. Since only isolates of the *C3* group seem to be adapted to *tetraploid* hosts, we suggest that this bottleneck could have been the result of a recent founder effect and/or a process of adaptation to *C. arabica* from a maladapted population. The most likely source of this introduction seems to be *Hv*'s isolates from diploid coffee hosts. Isolates infecting these hosts not only appear to be present for much longer, given their deep phylogenetic division, but they also share striking similarities in morphology and infection process with the *C3* isolates. Indeed, the fact that some *C2* and *C1* isolates are able to infect some *C. arabica* genotypes to a very limited extent would provide an entry point to this new ecological niche. The opportunity to shift to the new host would be given by the regular presence of tetraploids and *C. canephora* plants, either in wild or commercial plantations, within cruising range of one another (McCook and Vandermeer, 2015). Given the high genetic homogeneity of cultivated varieties of tetraploids (Anthony et al., 2002), the eventual adaptation of this population could have resulted in the inferred pathogen population boom. As mentioned above, this adaptation entails a fitness trade-off, since isolates of the *C3* group are no longer able to infect *C. canephora*. Coincidentally, this would also produce an effective pre-zygotic reproductive barrier known as immigrant inviability, where assortative mating arises as a by-product of host specialization (Giraud, 2006; Giraud et al., 2010). Nevertheless, the opportunity to hybridize could still present itself in non-commercial universally susceptible diploid coffee species. Interestingly,

a similar scenario has been tied to the emergence of several recent pathogens in modern agro-ecosystems (Geoghegan et al., 2016), including *Colletotrichum kahawae* causing Coffee Berry Disease (Silva et al., 2012a).

What remains surprising in this scenario is how the substantial loss of genetic diversity after the emergence of this new "domesticated" lineage is consistent with the rapid pace with which new rust pathotypes emerge. Despite the long-standing assumption that *Hv* is an asexual pathogen (Silva et al., 2006), our work provides several lines of evidence that recombination is occurring at least in the *C3* group. The presence of introgression between the *C3* and *C2* groups was the first hint pointing to the presence of recombination, since it is required for the backcrossing of genetic material from hybrids to the parental groups. Moreover, a very low proportion of SNP pairs were found to be under significant LD and the standardized index of association could not reject the null hypothesis of no linkage among markers. These metrics have been successfully used to uncover the existence of recombination in several fungal populations (Gladieux et al., 2011; Short et al., 2015), and their combination supports the existence of recombination of *Hv* isolates infecting tetraploids. Coincidentally, this would provide a viable mechanism for the quick generation of new allele combinations and, by extension, new pathotypes. The synergy of recombination, introgression and the ability of fungi to amplify favourable allele combinations through massive production of asexual spores, creates a plausible mechanism for pathogens to rapidly overcome resistant varieties, even from an initial pool of low genetic variation (Giraud et al., 2010). However, in the absence of functional sexual propagules, the mechanism by which recombination occurs still requires further investigation. Possible mechanisms include cryptosexuality within the urediniospores and parasexual nuclear recombination between two isolates after germ tube fusion or hyphal anastomosis, all of which can mimic the effects of sexual reproduction (Wang and McCallum, 2009; Carvalho et al., 2011; Vittal et al., 2012).

3.5.4 Conclusion and remarks on CLR disease management

Altogether, this work presents a striking example of how agro-ecosystems can have a dramatic impact on the population biology and evolution of plant pathogens. The discovery of three divergent genetic lineages within *Hv* with specialized pathogenic behaviour towards commercial coffee species and the ability to hybridize represents a paradigm shift in our current understanding of this pathogen's evolutionary history. It is clear that the taxonomic status of the three genetic lineages reported here needs to be revised in future studies. Referring to *Hv* as a single cohesive genetic unit is not only of little practical use in the future, but it can also be detrimental to quarantine practices. The introduction of isolates from the *C1* or *C2* groups into tetraploid coffee plantations does not carry the same risk as isolates from the *C3* group, and the reverse is true for *C. canephora* plantations. Mixing different lineages with the ability to hybridize has the potential of increasing the rate of pathotype emergence. Particularly greater care should be taken with nurseries of coffee germplasm and experimental stations, where multiple genotypes of different hybrids and species are stored nearby. Our results reveal that, if not adequately controlled and quarantined, isolates of *Hv* have strategies that make these locales ideal breeding grounds for the emergence of new hyper-virulent pathotypes. Subsequent studies of the *C3* group will be paramount to understand the frequency, direction and content that is transferred via introgression and recombination, which, ultimately, may be tightly linked to the capacity of the pathogen to overcome host resistance.

3.6 Supplementary data

3.6.1 Assembly strategy

3.6.1.1. Choice between PyRAD and Stacks

Both **PyRAD** and **Stacks** are established in the literature as pipelines developed for the assembly of RADseq data. However, they fundamentally differ in the way gaps are handled during a *de novo* assembly. While indels disrupt the clustering of homologous loci in **Stacks**, **PyRAD** relies on a clustering algorithm that accounts for the presence of gaps. Therefore, in data sets with a substantial amount of gaps, assemblies performed with **Stacks** should be more susceptible to errors that result from the presence of indels (such as frame-shifts, read drop-out), which could result in the elimination of reads or splitting of homologous sequences. To evaluate the impact that indels could have in our data set, we obtained the number and position of gaps for all assemblies using **PyRAD**. In average, 95% of the variable loci contained indels, though they often occurred at the end of the sequences (Table A.2.2). If we consider only indels occuring in the midle of the locus' sequence, an average of 25% of the variable loci contained gaps, that could produce gross assembly errors in **Stacks** due to frame-shifts. Given the substantial amount of indel found in our data, we proceeded our assemblies using only **PyRAD**.

3.6.1.2. Error rates

In order to optimize assembly parameters, several assemblies were performed in **PyRAD** with different parameters to assess their impact on four error metrics:

Locus error *Locus error* was divided into two sub-groups: *Total locus error* and *Partial locus error*. The *total locus error* refers to the proportion of loci from the full

sample that are missing for both replicates of a pair. *Partial locus error* refers to the number of loci that are absent from only one of the pair replicates.

Haplotype error Here, a haplotype is defined as the combination of all SNPs in a locus (RAD tag). If a locus contains three SNPs in replicate 1 of sample 1, there will only be an haplotype match between replicates if replicate 2 contains the exact same combination of SNPs. If there is a mismatch in one of the SNPs from the locus, it will score as an *haplotype error*.

SNP error In order to provide a more fine-grained resolution of mismatch between replicates, the *SNP error* quantifies the proportion of SNPs that differ between replicates. While the unit of the *haplotype error* rate was the entire locus, here the individual SNP represents the unit.

3.6.1.3. Optimization of assembly parameters

To investigate the impact of parameters on sequence assembly, three main parameters were tested with multiple values while fixing the remaining parameters across all tests. These parameters include the minimum read depth required to build a locus, the clustering threshold and the maximum shared heterozygosity. For each comparison, the final SNP data set was used without filtering and with a requirement of at least 50% taxa presence in the matrix.

Minimum depth Three values of minimum read depth were tested: 5, 10 and 15. In general, increasing the minimum read depth increased all error rates and decreased the number of final SNPs (Figure 3.9). The increase in error rates was higher for both *locus errors*, with an increase of 22.38% in *total error* rate and 5.42% in *partial error* rate. The error rates for haplotypes and SNPs only suffered marginal increases of 0.82% and 0.86%, respectively. In terms of number of loci, there was a linear decrease of total and variable loci with higher values of minimum read depth. As a result, the total number of SNPs also decreased linearly, from 19 801, with a minimum of 5 reads, to 12 600, with a minimum of 15 reads. While the increase of

locus error rates is expected due to the removal of low coverage loci, we did not expected the *haplotype* and *SNP error* rates to increase, even if only marginally, with the increased of minimum read coverage. In principle, variants in loci with higher read coverage should be called with higher certainty and therefore, we expected lower *SNP* and *haplotype errors*. However, a Kruskal-Wallis test revealed that these differences were not statistically significant for $p < 0.05$. Nevertheless, given the negative impact of increasing the minimum read coverage on the other rates, and the substantial reduction in the number of final SNPs, we used the value of 5 for the minimum depth parameter.

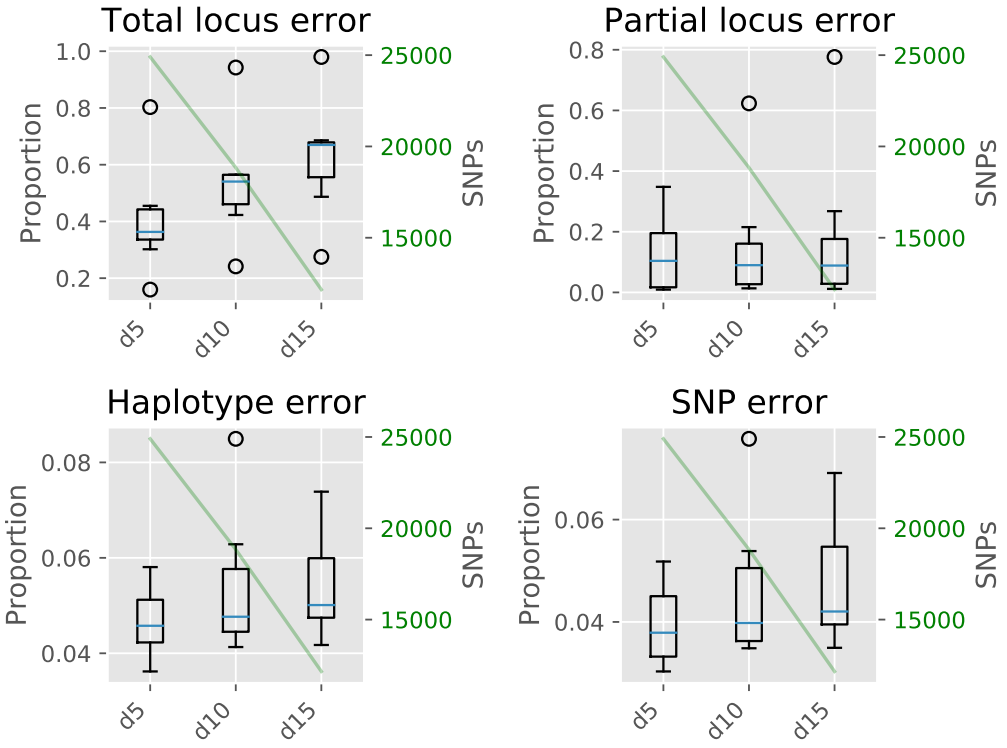

Figure 3.9. Effects of the minimum read depth parameter on the four error rates across assemblies. Minimum read depth is provided by the "d" value of the assembly name (*i.e.*, d15 means minimum read depth of 15).

Clustering threshold Five clustering threshold values were tested: 0.80, 0.85, 0.90, 0.95 and 0.97. Changing the clustering threshold had a small and non-significant effect on both *total* ($K = 2.23, p = 0.69$) and *partial* ($K = 1.11, p = 0.89$) error rates, with an average of 40.4% for *total locus* rate and 11.6% for *partial locus* error (Figure 3.10). However, decreasing this parameter had a significant impact on both *haplotype* ($K = 15.97, p = 0.003$) and *SNP* ($K = 24.65, p = 5.93e^{-05}$) error rates.

Concerning the number of loci, decreasing the clustering threshold also considerably decreased the number of total and variable loci, from 79 066 total and 28 898 variable, at a 0.97 clustering threshold, to 24 352 total and 9 917 variable loci, at a 0.80 clustering threshold. This represents a decrease in 69.20% of total loci and 59.28% of variable loci. Likewise, the number of total SNPs drops from 19 801, at a 0.97 clustering threshold, to 8 540, at a 0.80 clustering threshold, representing a 56.87% decrease. Given these results, we selected the optimum clustering threshold of 0.97 for the final assembly.

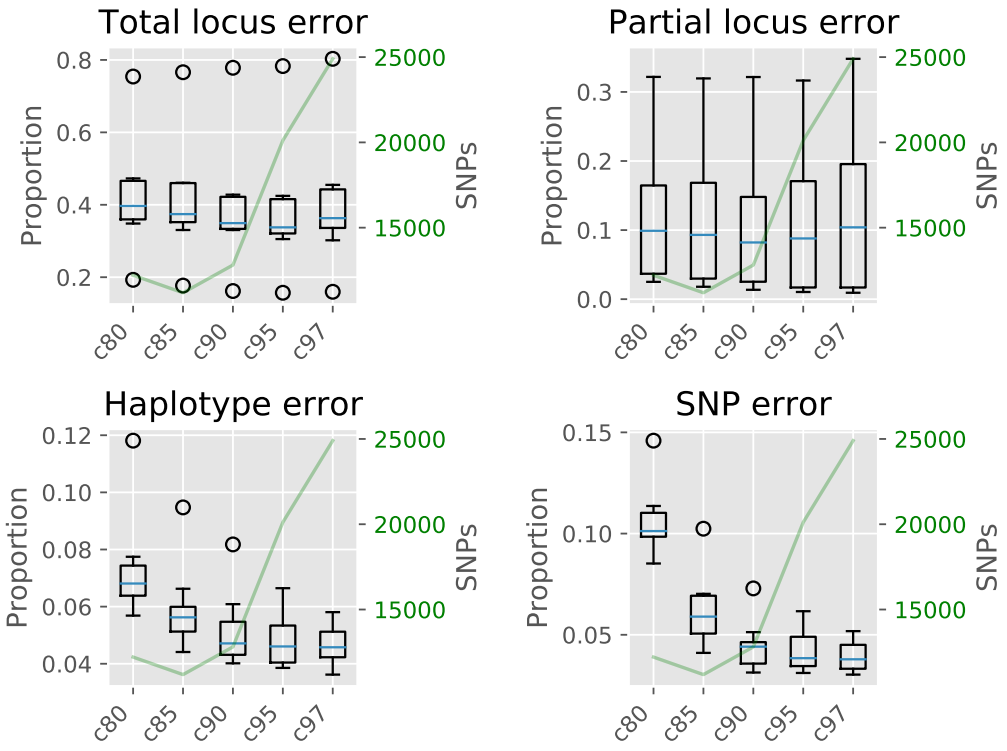

Figure 3.10. Effects of the clustering threshold parameter on the four error rates across assemblies. Clustering threshold is provided by the "c" value of the assembly name (*i.e.*, c80 means clustering threshold of 0.80).

Maximum shared heterozigosity Three values of maximum shared heterozigosity were tested: 2, 5 and 10. Increasing the value of this parameter had a small and non-significant impact on the *total* ($K = 0.62, p = 0.73$) and *partial* ($K = 0.06, p = 0.97$) error rates, with an average of 40.21% and 11.70% error rates, respectively (Figure 3.11). Considering the *haplotype* ($K = 5.14, p = 0.077$) and *SNP* ($K = 4.82, p = 0.090$) error rates, even though there were no statistically significant differences between parameter values, there was a substantial increase in error rate variance with higher

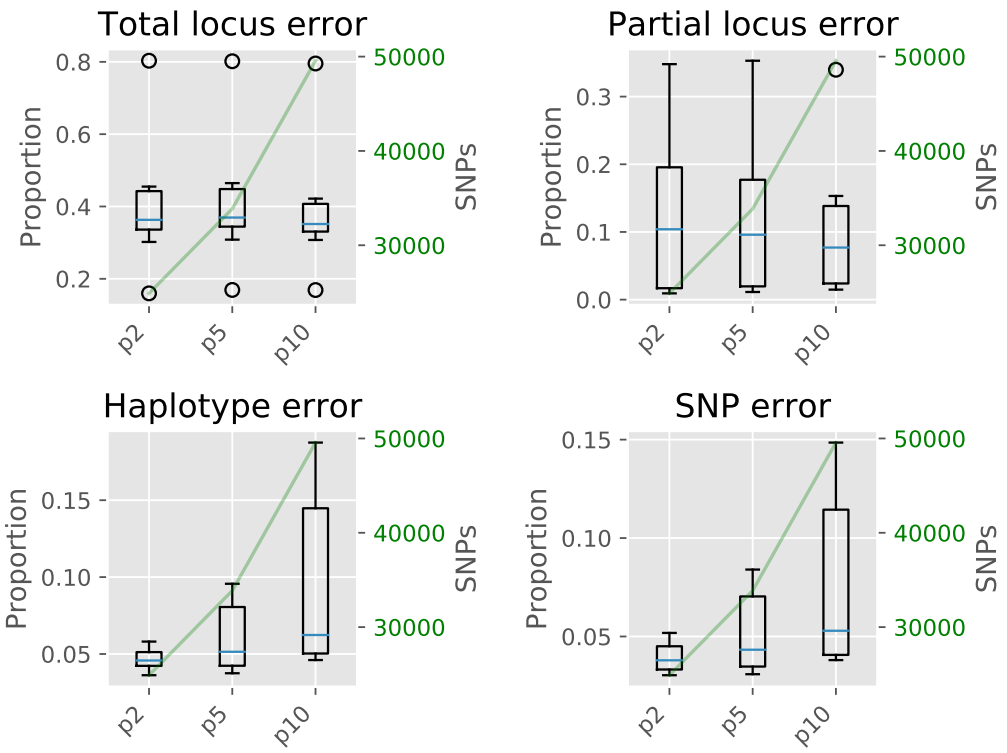

Figure 3.11. Effects of the maximum shared heterozygosity parameter on the four error rates across assemblies. Maximum shared heterozygosity is provided by the "p" value of the assembly name (*i.e.*, p2 means maximum shared heterozygosity of 2).

parameter values. This increase in variance was mainly due to error rate increases in replicates from diploid hosts and was also accompanied by an increase in both total and variable loci, as well as in the total number of SNPs. The total number of SNPs rose from 19 801, with a parameter value of 2, to 35 908, with a parameter value of 10, representing an increase of 44.86%. Even though *haplotype* and *SNP* error rates were not significantly higher for higher parameter values, we were wary of the increase in error rate variance and decided to use the parameter value of 2, which yielded the smallest error rates. Nevertheless, the main analysis of the paper were also conducted with the assembly that specified the maximum shared heterozigosity value of 10, to assess the robustness of our findings when using a data set with a higher amount of SNPs. We found that the results remained qualitatively equal and quantitatively similar.

3.7 References

Agapow, P. and A Burt (2001). „Indices of multilocus linkage disequilibrium“. *Molecular Ecology Notes* 1.1, pp. 101–102.

Anderson, T., B. VonHoldt, S. Candille, M Musiani, C Greco, D. Stahler, D. Smith, B Padhukasahasram, E Randi, J. Leonard, et al. (2009). „Molecular and evolutionary history of melanism in North American Gray Wolves“. *Science* 323, pp. 1339–1343.

Anthony, F., M. C. Combes, C. Astorga, B. Bertrand, G. Graziosi, and P. Lashermes (2002). „The origin of cultivated Coffea arabica L. varieties revealed by AFLP and SSR markers“. *Theoretical and Applied Genetics* 104.5, pp. 894–900.

Avelino, J., M. Cristancho, S. Georgiou, P. Imbach, L. Aguilar, G. Bornemann, P. Läderach, F. Anzueto, A. J. Hruska, and C. Morales (2015). „The coffee rust crises in Colombia and Central America (2008–2013): impacts, plausible causes and proposed solutions“. *Food Security* 7.2, pp. 303–321.

Bennett, C., M. C. Aime, and G. Newcombe (2011). „Molecular and pathogenic variation within Melampsora on Salix in western North America reveals numerous cryptic species.“ *Mycologia* 103.5, pp. 1004–1018.

Bettencourt, A. and C. J. Rodrigues (1988). „Principles and practice of coffee breeding for resistance to rust and other diseases“. In: *Coffee Agronomy, Vol. 4*. Ed. by R. Clarke and R Macrae. Lodon and New York: Elsevier, pp. 199–234.

Cabral, P. G. C., E. Maciel-Zambolim, S. A. S. Oliveira, E. T. Caixeta, and L. Zambolim (2016). „Genetic diversity and structure of Hemileia vastatrix populations on Coffea spp.“ *Plant Pathology* 65.2, pp. 196–204.

Carvalho, C. R., R. C. Fernandes, G. M. A. Carvalho, R. W. Barreto, and H. C. Evans (2011). „Cryptosexuality and the Genetic Diversity Paradox in Coffee Rust, Hemileia vastatrix“. *PLoS ONE* 6.11, e26387.

Catchen, J., P. a. Hohenlohe, S. Bassham, A. Amores, and W. a. Cresko (2013). „Stacks: an analysis tool set for population genomics“. *Molecular Ecology* 22.11, pp. 3124–3140.

Cenci, A., M. C. Combes, and P. Lashermes (2012). „Genome evolution in diploid and tetraploid Coffea species as revealed by comparative analysis of orthologous genome segments“. *Plant Molecular Biology* 78.1-2, pp. 135–145.

Cristancho, M. A., D. O. Botero-Rozo, W. Giraldo, J. Tabima, D. M. Riaño-Pachón, C. Escobar, Y. Rozo, L. F. Rivera, A. Durán, S. Restrepo, et al. (2014). „Annotation of a hybrid partial genome of the coffee rust (Hemileia vastatrix) contributes to the gene repertoire catalog of the Pucciniales.“ *Frontiers in Plant Science* 5, p. 594.

Danecek, P., A. Auton, G. Abecasis, C. a. Albers, E. Banks, M. a. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, et al. (2011). „The variant call format and VCFtools“. *Bioinformatics* 27.15, pp. 2156–2158.

D'Oliveira, B (1954). „As ferrugens do cafeeiro“. *Rev. Cafe Port.* 1, pp. 5–13.

Drenth, A. and D. I. Guest (2016). „Fungal and Oomycete Diseases of Tropical Tree Fruit Crops“. *Annual Review of Phytopathology* 54.1, pp. 373–395.

Drummond, A. J., M. a.D Suchard, D. Xie, and A. Rambaut (2012). „Bayesian phylogenetics with BEAUti and the BEAST 1.7“. *Molecular Biology and Evolution* 29.8, pp. 1969–1973.

Eaton, D. and R. Ree (2013). „Inferring Phylogeny and Introgression using RADseq Data : An Example from Flowering Plants (Pedicularis : Orobanchaceae)“. *Systematic Biology* 62.5, pp. 689–706.

Eaton, D. a. R. (2014). „PyRAD: assembly of de novo RADseq loci for phylogenetic analyses“. *Bioinformatics* 30.13, pp. 1844–1849.

Eriksson, A. and A. Manica (2012). „Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins“. *Proceedings of the National Academy of Sciences of the United States of America* 109.35, pp. 13956–13960.

Etter, P., S Bassham, P. Hohenlohe, E. Johnson, and W. Cresko (2011). „SNP Discovery and Genotyping for Evolutionary Genetics using RAD Sequencing“. In: *Molecular Methods for Evolutionary Genetics*. Ed. by V Orgogozo and M. Rockman. Vol. 772. Methods in Molecular Biology. Totowa, NJ: Humana Press, pp. 157–178.

Flor, H. (1955). „Host-parasite interactions in flax rust: its genetics and other implications“. *Phytopathology* 45, pp. 680–685.

Gandon, S., T. Day, C. J. E. Metcalf, and B. T. Grenfell (2016). „Forecasting Epidemiological and Evolutionary Dynamics of Infectious Diseases“. *Trends in Ecology & Evolution* 31, pp. 776–788.

Geoghegan, J. L., A. M. Senior, and E. C. Holmes (2016). „Pathogen population bottlenecks and adaptive landscapes: overcoming the barriers to disease emergence“. *Proceedings of the Royal Society B: Biological Sciences* 283.1837, p. 20160727.

Giraud, T (2006). „Selection against migrant pathogens: the immigrant inviability barrier in pathogens.“ *Heredity* 97, pp. 316–318.

Giraud, T., P. Gladieux, and S. Gavrilets (2010). „Linking the emergence of fungal plant diseases with ecological speciation.“ *Trends in Ecology & Evolution* 25.7, pp. 387–395.

Gladieux, P., E. Vercken, M. C. Fontaine, M. E. Hood, O. Jonot, A. Couloux, and T. Giraud (2011). „Maintenance of Fungal Pathogen Species That Are Specialized to Different Hosts: Allopatric Divergence and Introgression through Secondary Contact.“ *Molecular Biology and Evolution* 28.1, pp. 459–471.

Gouveia, M. M. C., A. Ribeiro, V. M. P. Várzea, and C. J. Rodrigues (2005). „Genetic diversity in Hemileia vastatrix based on RAPD markers.“ *Mycologia* 97.2, pp. 396–404.

Grünwald, N., B. A. McDonald, and M. G. Milgroom (2016). „Population Genomics of Fungal and Oomycete Pathogens“. *Annual Review of Phytopathology* 54.1, pp. 323–346.

Harrison, R. G. and E. L. Larson (2014). „Hybridization, introgression, and the nature of species boundaries“. *Journal of Heredity* 105.S1, pp. 795–809.

Heled, J. and A. J. Drummond (2008). „Bayesian inference of population size history from multiple loci“. *BMC Evolutionary Biology* 8.1, p. 289.

Kamvar, Z. N., J. F. Tabima, and N. J. Grünwald (2014). „Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction.“ *PeerJ* 2, e281.

Kolmer, J., J. Liu, and M. Sies (1995). „Virulence and molecular polymorphism in Puccinia recondita f. sp. tritici in Canada“. *Phytopathology* 85, pp. 276–285.

Maia, T. a., E. Maciel-Zambolim, E. T. Caixeta, E. S. G. Mizubuti, and L. Zambolim (2013). „The population structure of Hemileia vastatrix in Brazil inferred from AFLP“. *Australasian Plant Pathology* 42.5, pp. 533–542.

Mastretta-Yanes, A, N Arrigo, N Alvarez, T. H. Jorgensen, D Piñero, and B. C. Emerson (2015). „Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference“. *Molecular Ecology Resources* 15.1, pp. 28–41.

McCook, S. and J. Vandermeer (2015). „The Big Rust and the Red Queen: Long-Term Perspectives on Coffee Rust Research.“ *Phytopathology* 105.9, pp. 1164–1173.

McDonald, B. A. and C. Linde (2002). „Pathogen population genetics, evolutionary potential, and durable resistance“. *Annual Review of Phytopathology* 40.1, pp. 349–379.

Menardo, F., C. R. Praz, S. Wyder, R. Ben-David, S. Bourras, H. Matsumae, K. E. McNally, F. Parlange, A. Riba, S. Roffler, et al. (2016). „Hybridization of powdery mildew strains gives rise to pathogens on novel agricultural crop species“. *Nature Genetics* 48.2, pp. 201–205.

Milgroom, M. G., M. D. M. Jiménez-Gasco, C. Olivares García, M. T. Drott, and R. M. Jiménez-Díaz (2014). „Recombination between Clonal Lineages of the Asexual Fungus Verticillium dahliae Detected by Genotyping by Sequencing“. *PLoS ONE* 9.9, e106740.

Miller, M., W. Pfeiffer, and T. Schwartz (2010). „Creating the CIPRES Science Gateway for inference of large phylogenetic trees“. In: *Proceedings of the Gateway Computing Environments Workshop (GCE)*, pp. 1–8.

Nunes, C. C., L. A. Maffia, E. S. G. Mizubuti, S. H. Brommonschenkel, and J. C. Silva (2009). „Genetic diversity of populations of Hemileia vastatrix from organic and conventional coffee plantations in Brazil“. *Australian Plant Pathology* 38, pp. 445–452.

Pina-Martins, F., D. N. Silva, J. Fino, and O. S. Paulo (2017). „Structure_threader : An improved method for automation and parallelization of programs structure , fastStructure and MavericK on multicore CPU systems“. *Molecular Ecology Resources* 17.6, e268–e274.

Prakash, N., J Devasia, K Das Divya, B. Manjunatha, H. Seetharam, A Kumar, and Jayarama (2014). „Breeding for rust resistance in Arabica – where we are and what next?“ In: *Proceedings of the 25th International Conference on Coffee Science (ASIC)*, B10.

Pringle, a, D. M. Baker, J. L. Platt, J. P. Wares, J. P. Latgé, and J. W. Taylor (2005). „Cryptic speciation in the cosmopolitan and clonal human pathogenic fungus Aspergillus fumigatus.“ *Evolution* 59.9, pp. 1886–1899.

Raj, A., M. Stephens, and J. K. Pritchard (2014). „fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Datasets.“ *Genetics* 197, pp. 573–589.

Rambaut, A, M. Suchard, D Xie, and A. Drummond (2014). *Tracer v1.6, Available from http://beast.bio.ed.ac.uk/Tracer*.

Rheindt, F. E., M. K. Fujita, P. R. Wilton, and S. V. Edwards (2014). „Introgression and phenotypic assimilation in zimmerius flycatchers (Tyrannidae): Population genetic and phylogenetic inferences from genome-wide SNPs“. *Systematic Biology* 63.2, pp. 134–152.

Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. a. Suchard, and J. P. Huelsenbeck (2012). „MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space.“ *Systematic biology* 61.3, pp. 539–542.

Rozo, Y., C. Escobar, Á. Gaitán, and M. Cristancho (2012). „Aggressiveness and Genetic Diversity of Hemileia vastatrix During an Epidemic in Colombia“. *Journal of Phytopathology* 160.11-12, pp. 732–740.

Short, D. P. G., S. Gurung, P. Gladieux, P. Inderbitzin, Z. K. Atallah, F. Nigro, G. Li, S. Benlioglu, and K. V. Subbarao (2015). „Globally invading populations of the fungal plant pathogen Verticillium dahliae are dominated by multiple divergent lineages“. *Environmental Microbiology* 17.8, pp. 2824–2840.

Silva, D. N., P. Talhinhas, L. Cai, L. Manuel, E. K. Gichuru, A. Loureiro, V. Várzea, O. S. Paulo, and D. Batista (2012a). „Host-jump drives rapid and recent ecological speciation of the emergent fungal pathogen Colletotrichum kahawae.“ *Molecular Ecology* 21.11, pp. 2655–2670.

Silva, D. N., P. Talhinhas, V. Várzea, L. Cai, O. S. Paulo, and D. Batista (2012b). „Application of the Apn2/MAT locus to improve the systematics of the Colletotrichum gloeosporioides complex: an example from coffee (Coffea spp.) hosts.“ *Mycologia* 104.2, pp. 396–409.

Silva, M. C., L. Guerra-Guimarães, A. Loureiro, and M. R. Nicole (2008). „Involvement of peroxidases in the coffee resistance to orange rust (Hemileia vastatrix)“. *Physiological and Molecular Plant Pathology* 72.1-3, pp. 29–38.

Silva, M., V Várzea, L Guerra-Guimarães, H Azinheira, D Fernandez, A.-S. Petitot, B Bertrand, P Lashermes, and M Nicole (2006). „Coffee resistance to the main diseases: leaf rust and coffee berry disease“. *Brazilian Journal of Plant Physiology* 18, pp. 119–147.

Stamatakis, A. (2014). „RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies“. *Bioinformatics* 30.9, pp. 1312–1313.

Staubach, F., A. Lorenc, P. W. Messer, K. Tang, D. a. Petrov, and D. Tautz (2012). „Genome Patterns of Selection and Introgression of Haplotypes in Natural Populations of the House Mouse (Mus musculus)“. *PLoS Genetics* 8.8. Ed. by M. H. Kohn, e1002891.

Stukenbrock, E. H. (2016). „Hybridization speeds up the emergence and evolution of a new pathogen species“. *Nature Genetics* 48.2, pp. 113–115.

Stukenbrock, E. H. and T. Bataillon (2012). „A Population Genomics Perspective on the Emergence and Adaptation of New Plant Pathogens in Agro-Ecosystems“. *PLoS Pathogens* 8.9, e1002893.

Stukenbrock, E. H. and B. A. McDonald (2008). „The origins of plant pathogens in agro-ecosystems.“ *Annual Review of Phytopathology* 46, pp. 75–100.

Stukenbrock, E. H., T. Bataillon, J. Y. Dutheil, T. T. Hansen, R. Li, M. Zala, B. a. McDonald, J. Wang, and M. H. Schierup (2011). „The making of a new pathogen: insights from comparative population genomics of the domesticated wheat pathogen Mycosphaerella graminicola and its wild sister species.“ *Genome Research* 21.12, pp. 2157–2166.

Stukenbrock, E. H. (2013). „Evolution, selection and isolation: a genomic view of speciation in fungal plant pathogens“. *New Phytologist* 199, pp. 895–907.

Talas, F. and B. A. McDonald (2015). „Genome-wide analysis of Fusarium graminearum field populations reveals hotspots of recombination.“ *BMC Genomics* 16.1, p. 996.

Talhinhas, P., H. G. Azinheira, B. Vieira, A. Loureiro, S. Tavares, D. Batista, E. Morin, A.-S. Petitot, O. S. Paulo, J. Poulain, et al. (2014). „Overview of the functional virulent genome of the coffee leaf

rust pathogen Hemileia vastatrix with an emphasis on early stages of infection“. *Frontiers in Plant Science* 5, p. 88.

Talhinhas, P., D. Batista, I Diniz, A Vieira, D. Silva, A Loureiro, S Tavares, A. Pereira, H. Azinheira, L Guerra-Guimarães, et al. (2017). „Pathogen profile The coffee leaf rust pathogen Hemileia vastatrix : one and a half centuries around the tropics“. *Molecular Plant Pathology* 18.8, pp. 1039–1051.

Todesco, M., M. A. Pascual, G. L. Owens, K. L. Ostevik, B. T. Moyers, S. Hübner, S. M. Heredia, M. A. Hahn, C. Caseys, D. G. Bock, et al. (2016). „Hybridization and extinction“. *Evolutionary Applications* 9.7, pp. 892–908.

Twyford, a. D. and R. a. Ennos (2012). „Next-generation hybridization and introgression“. *Heredity* 108.3, pp. 179–189.

Vittal, R., H. C. Yang, and G. L. Hartman (2012). „Anastomosis of germ tubes and migration of nuclei in germ tube networks of the soybean rust pathogen, Phakopsora pachyrhizi“. *European Journal of Plant Pathology* 132.2, pp. 163–167.

Wang, X. and B. McCallum (2009). „Fusion body formation, germ tube anastomosis, and nuclear migration during the germination of urediniospores of the wheat leaf rust fungus, Puccinia triticina.“ *Phytopathology* 99.12, pp. 1355–64.

Zambolim, L. (2016). „Current status and management of coffee leaf rust in Brazil“. *Tropical Plant Pathology* 41.1, pp. 1–8.

Zhan, J., P. H. Thrall, J. Papaïx, L. Xie, and J. J. Burdon (2015). „Playing on a Pathogen's Weakness: Using Evolution to Guide Sustainable Plant Disease Control Strategies“. *Annual Review of Phytopathology* 53.1, pp. 19–43.

Zhao, P., Q. H. Wang, C. M. Tian, and M. Kakishima (2015). „Integrating a numerical taxonomic method and molecular phylogeny for species delimitation of melampsora species (melampsoraceae, pucciniales) on willows in China“. *PLoS ONE* 10.12, pp. 1–18.

Zheng, X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie, and B. S. Weir (2012). „A high-performance computing toolset for relatedness and principal component analysis of SNP data“. *Bioinformatics* 28.24, pp. 3326–3328.

4

TriFusion: Streamlining phylogenomic data gathering, processing and visualization

Diogo N. Silva [1,2,3], Fernando Alves [4], Dora Batista [1,2,3], Octávio S. Paulo [1]

[1] Centro de Investigação das Ferrugens do Cafeeiro, Instituto Superior de Agronomia, Universidade de Lisboa, Oeiras, Portugal.

[2] Computational Biology and Population Genomics group, cE3c – Centre for Centre for Ecology Evolution and Environmental Changes, Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal.

[3] Departamento de Biologia e CESAM - Centro de Estudos do Ambiente e do Mar, Universidade de Aveiro, Aveiro, Portugal.

[4] LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal

4.1 Abstract

Phylogenomic studies are now dealing with massive data sets that present a considerable technical challenge to gather, process and visualize. These tasks usually require substantial bioinformatic expertise and/or server-grade computers to handle the amount of data and overcome the performance limitations of the available tools. In this paper we present TriFusion, a feature rich application capable of gathering, processing and visualizing phylogenomic data that blends high performance and efficiency with user-friendly graphical and command line interfaces. The main goal of TriFusion is to streamline the laborious and complex tasks that are required before

starting the actual phylogenetic analyses, from the assembly to the manipulation and exploration of sequence alignment data. This is achieved via three main modules: The **Orthology** module allows the identification, exploration and filtering of ortholog genes across multiple species using an adaptation of the popular OrthoMCL framework; The **Process** module offers a wide array of options for converting, concatenating or manipulating up to thousands of sequence alignment files with high flexibility and efficiency. Even complex and intensive tasks, such as a filtering alignments and/or alignment columns, collapsing sequences, coding insertion-deletion patterns, handling partitions and substitution models can be easily executed for large data sets; The **Statistics** module offers a myriad of graphical and statistical analyses that allow the effortless visualization of sequence alignment data and the detection of alignment/taxon outliers according to missing data, sequence variation and/or sequence size. Altogether, TriFusion offers a novel and comprehensive suite of operations that can greatly facilitate several stages of any phylogenomic study and provides unparalleled performance when processing massive data sets for downstream phylogenetic analyses. The software is open source, cross-platform, easy to install and freely available at http://odiogosilva.github.io/TriFusion.

4.2 Introduction

In the last few years, the striking advances in high-throughput sequencing technologies have dramatically increased and driven the scope and ambition of phylogenetic studies into the realm of phylogenomics. However, the wealth of data that is currently being generated also presents an exceptional technical challenge. Phylogenomic pipelines have become very complex, requiring the execution of several laborious and non-trivial steps concerning the assembly, manipulation and interpretation of the sequence data before starting the actual phylogenetic analyses. Indeed, these steps represent the bulk of operations that researches have to perform in a typical phylogenomics study. In the initial stage of assembling sequence data, the identification of orthologs across multiple genomes is a frequent approach for which there are several software tools available, such as Ortholog-Finder (Horiike et al., 2016), OrthoFinder (Emms and Kelly, 2015) or OrthoMCL (Li and Stephens, 2003).

These are powerful tools that have been widely used in phylogenomic studies, but also have a strong reliance on third-party software, thus requiring a setup process that non-bioinformaticians may find hard and tedious to follow. Concerning the manipulation of data matrices for downstream phylogenetic analyses, the most common tasks include the conversion, concatenation and filtering of alignment files. Many programs and services have been developed to aid researchers with these tasks (Roure et al., 2007; Vaidya et al., 2011; Lischer and Excoffier, 2012; Kück and Longo, 2014; Bond et al., 2017), albeit focusing on different aspects of data processing and with varying degrees of accessibility to end-users. For example, PGDSpider (Lischer and Excoffier, 2012) has a sizable list of input and output formats for conversion but cannot be used for concatenation of multiple files. BuddySuite (Bond et al., 2017) provides a staggering amount of manipulation options but can only be used through a command line, which may deter users not familiar with such interface or with scripting languages. However, all these tools have severe performance limitations when dealing with large scale data sets generated from current phylogenomic projects. In fact, some data sets cannot be processed at all with the currently available tools. Consequently, there is a need for powerful tools that are able to cope and scale with large phylogenomic data sets while providing a user-friendly interface that significantly opens phylogenomic approaches to a wider audience.

4.3 Description

This paper presents TriFusion, a feature rich Python application capable of gathering, processing and visualizing phylogenomic data that blends high performance and efficiency with a user-friendly interface design. Features are available via a graphical user interface (GUI) or via command line (CLI), and are divided into three modules representing three important stages in phylogenomic projects: **Orthology**, allows the identification and exploration of ortholog genes across multiple proteomes; **Process**, allows the conversion, concatenation, filtering, collapsing, gap coding, consensus creation and other manipulations on alignment files; **Statistics**, provides dozens of graphical and statistical operations meant to visualize and bet-

ter understand the data contained in multiple alignment files (Figure 4.1). For a detailed feature description, please consult the user guide or TriFusion's website (http://odiogosilva.github.io/TriFusion/content/features.html).

4.3.1 Orthology: Search and explore orthologs

The **Orthology** module was designed to streamline the task of detecting and exploring ortholog groups across multiple proteomes and is divided into two operations: Search and Explore. In the *Search* operation, we adapted the framework of the popular software OrthoMCL (Li and Stephens, 2003) to detect ortholog groups, introducing a few key modifications aimed at increasing performance and user-accessibility. First, by porting the original Perl and MySQL code to Python and SQLite, TriFusion executes the pipeline with fewer dependencies and using a self-contained database that precludes the setup of a MySQL server by the user. Second, we incorporated **USEARCH** (Edgar, 2010) instead of NCBI **BLAST+** (Camacho et al., 2009) into the pipeline, since it executes the all-vs-all protein search operation orders of magnitude faster and with comparable sensitivity. Third, binary executables of the required **MCL** program (Enright et al., 2002) come bundled in TriFusion for all major operating systems, allowing the pipeline to be fully cross-platform. Finally, the entire pipeline can be executed automatically with multiple **MCL** inflation values specified in the same run, generating protein sequence files for each ortholog group in addition to the usual output of OrthoMCL. All original options of OrthoMCL are accessible and TriFusion includes additional ones to filter ortholog groups according to the maximum number of genes copies (gene filter) and minimum number of taxa (taxa filter). Therefore, users can take full advantage of the popular OrthoMCL framework using a new interface that is much more user-friendly; the processing framework requires a minimal setup, is faster and produces filtered output files ready for downstream analyses.

The *Explore* operation allows users to load the results from one or more orthology search operations and interactively explore the data with the aid of filters and visualization tools. When the data is loaded, TriFusion provides immediate visual feedback with information on the composition of the results, including the number

of proteins, taxa and ortholog groups (both with and without filters). A common approach with the OrthoMCL pipeline is to detect ortholog groups using different inflation values for the MCL algorithm. In TriFusion, the results of different searches can be easily compared in order to understand the impact that the inflation parameter has on the final number of orthologs, while adjusting the gene and taxa filters. Individual group files can be further explored with several graphical visualization analyses that allow users to gain insights on the distribution of taxa across ortholog groups, the missing data across taxa or the distribution of gene copies per taxa or ortholog groups. Plots can be visualized inside the application and seamlessly updated when the gene and taxa filter values are changed and/or when taxa are temporarily removed. Alternatively, a full report with all available graphical visualization analyses can be easily obtained, in HTML format. At any point during the data exploration, the ortholog groups passing the current filters can be exported as protein and/or nucleotide sequence files. This is an important feature that generates the ortholog data used for downstream analyses, which is missing in the OrthoMCL package scripts. To export ortholog groups into protein sequences, only a protein database file is required, which is already generated by TriFusion when performing a Search operation. To export into nucleotide sequences, in addition to the protein database file, users must also provide the CDS or transcript files containing the corresponding nucleotide sequences of the input proteome files. These files are usually generated alongside the proteome files in genome sequencing projects. Since TriFusion uses USEARCH to find nucleotide sequences in the CDS/transcript files that have a perfect match in the protein database file, there is no requirement for matching Fasta headers between files and users can be certain that the converted nucleotide sequences are retrieved correctly.

4.3.2 Process: Manipulation and processing of alignments

The **Process** module deals with the conversion, concatenation and general manipulation of alignment files using highly efficient data structures and techniques that allow the fast processing of massive data sets ($\geq$ 50k files or files with $\geq$ 1Gb) in regular desktop computers (See section 4.4 for algorithm design and implementation). Besides efficiency and performance, this module was designed to be easy to

use and to provide a diverse set of intuitive and powerful options that allow the swift manipulation of alignment data matrices for downstream analyses. Protein and/or nucleotide alignment files in one or more popular formats (Fasta, Nexus, Phylip, Stockholm and **PyRAD**'s loci format from RADseq data), can be quickly and simultaneously loaded into **TriFusion** using several available methods. Input formats, sequence types, partitions schemes, substitution models (for Nexus format) and other relevant information are automatically detected and retrieved from each file, regardless of its extension. At any time, custom partition schemes and substitution models can be provided via a text file in Nexus or **RAxML**'s partition formats and/or modified directly within the application. Summary information of the loaded data is readily available in the application's side panel, which includes tabs for files, taxa and partitions, along with more detailed information for each file/taxon/partition. In this side panel, users will also find several convenient options and features that allow them to easily modify the "active" data set, create custom data set groups, save the current session for quick loading in future sessions or view the list/queue of currently active tasks. Since alignment manipulations are so frequently performed in phylogenomics/phylogenetics, these features were implemented to reduce the burden of repetitive tasks and to increase the agility of users when processing their data sets. For instance, after defining taxa groups for several taxonomic groups of interest in the data set and/or alignment groups to separate mitochondrial from nuclear alignments, it is possible to perform the same or entirely different tasks on different sets of files and taxa, within the same session, by simply changing a name in a drop-down menu.

As shown in Figure 4.1, data processing is initiated by selecting one of three main operations: Conversion, where each individual alignment file is converted; Concatenation, where alignment files are merged together according to the scheme that is appropriate for each specified output format; or Reverse concatenation, where one or more alignment files can be exported as individual alignment files according to a user-specified partition scheme. Output alignments can be converted/concatenated into one or more popular formats in phylogenomics. Depending on the selected main operation, different output formats become available. For instance, formats such as **SNAPP** or **MCMCTree**'s depend on the existence of partitions and can

only be selected using the Concatenation operation. In either case, TriFusion will execute the conversion/concatenation of the sequence data automatically handling data sets with missing taxa for some alignments, mixtures of nucleotide, protein and/or binary data, custom partition sets, custom substitution models and auxiliary alignment weight files such as those produced by ZORRO (Wu et al., 2012). Indeed, a particularly attractive feature of TriFusion is that even these formatting issues and details that often require manual intervention or scripting knowledge are seamlessly handled.

In addition to the main operation, TriFusion offers a wide range of secondary operations that can further manipulate the data. One of the most important and frequently used manipulations in phylogenomic data sets is the filtering of alignment files and/or alignment columns according to some metric. TriFusion allows filters to be set according to taxa groups (e.g. sets of alignments that maximize the data for a taxonomic group, or that exclude a group of taxa), codon position, missing data and/or sequence variation. Whenever filters are active, a final report is also generated informing the user of how many alignments were filtered and by which filter. Additional operations include the collapse of identical sequences to create an alignment with only unique sequences, the creation of consensus sequences for each alignment with several options for handling sequence variation, or the coding of indel patterns as a binary state matrix at the end of the alignment. These operations are all optional and can be specified either individually or in combination to perform intensive and complex procedures. It is worth noting that all secondary operations may have an impact on the final partition schemes, alignment lengths and types of data. For instance, filtering alignments removes entire partitions, filtering alignment columns modifies the lengths of alignments and partition ranges, and coding insertion-deletion patterns introduces a new partition with a different type of sequence data. Once again, all these modifications are taken into account and automatically handled by TriFusion so that the output files reflect these changes and are correctly formatted for downstream analyses. In the end, the output of these processing procedures is ready for a wide range of analyses, from general phylogenetic inference to functional annotation.

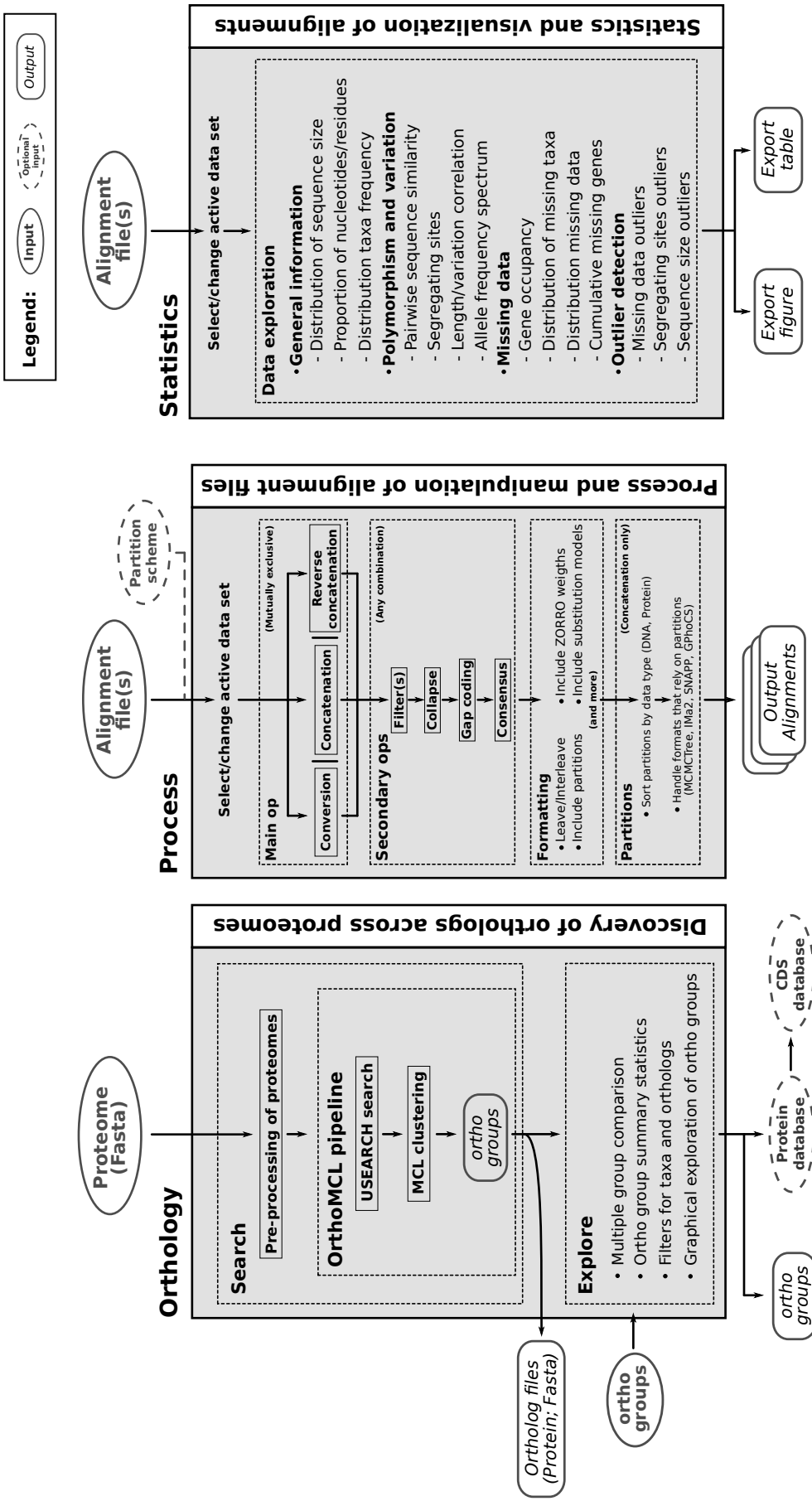

Figure 4.1. Diagram representing TriFusion's workflow across its three main modules.

4.3.3 Statistics: Statistics and visualization of alignment data

One of the challenges that researchers face when dealing with large data sets of alignment files is how to analyse the characteristics and composition of their data. The **Statistics** module addresses that challenge by providing a set of graphic and statistical analyses that allow users to effortlessly explore and visualize different facets of their alignment data. Although they can be used in data sets of any size, these analyses were optimized for large alignment data sets, which are also the most difficult to explore in an efficient and timely fashion. After loading alignment data into TriFusion, as the user navigates to the **Statistics** module the first task it will undertake is the calculation of several summary statistics focused on missing data and sequence variation. These are displayed as an overview board for the global active data set or as a gene table where the summary statistics are discriminated for each gene. To continue exploring the data, numerous options are available and sorted into four thematic categories that focus on particular aspects of the data: General Information, Polymorphism and Variation, Missing data and Outlier Detection. Moreover, for the majority of the plotting options there are up to three available plot types that provide different perspectives on the same analysis: single gene, average and per species. Single gene plot types explore a given metric along the length of a single alignment; average plot types focus on the distribution of a given metric across the entire alignment data set; and per species plot types focus on the distribution of a metric for each taxon/species. All plots are displayed within the application and almost all can be exported as a figure and as a table in CSV format. For convenience, users can create multiple taxa and file groups, as in the **Process** module, and then switch between them to update the plots with the new active data set. Since users may frequently switch between analyses and plot types during the data exploration, we also implemented a "fast-switch" method where the analysis and plot generation for any given option is calculated only once and then stored locally as a temporary file.

It is worth noting that the Outlier Detection category is particularly useful for phylogenomic data sets, since it allows users to detect the presence of outlier genes and/or

taxa according to sequence length, sequence variation and missing data. TriFusion uses the Median Absolute Deviation (MAD) method to find outliers with modified Z-scores that ensure that the MAD provides a good estimate independently of the sample size (Iglewicz and Hoaglin, 1993). These analyses can provide insights on the presence of particular alignments or taxa that may show an abnormal behavior according to any of the mentioned metrics and, thus, identify potential problems from upstream analyses before undertaking the actual phylogenetic analyses.

4.3.4 Command line versions

The majority of TriFusion's features are also available through command line programs that can be executed in headless servers and easily integrated in automatic pipelines. The **Orthology** search operation is implemented in orthomcl_pipeline, the **Process** module is implemented in TriSeq and the **Statistics** module is implemented in TriStat.

4.4 Algorithm description and performance

The development of TriFusion as a modern desktop application with GUI and CLI interfaces was a substantial undertaking that included activities that went from the structuring of the package, to the design and implementation of backend algorithms, the development of a graphical user interface, setup of a comprehensive suite of tests, packaging and distribution, and the creation of a website and tutorials. In this section I will describe some of the technologies, algorithms and optimizations used by TriFusion's **Process** module backend that allow it to process massive alignment data sets quickly and efficiently.

TriFusion's code was mainly written in Python2 and most of the implementations described in this section make use of the SQLite database engine.

4.4.1 Alignment parsers

The parsing of alignment files is handled by the `Alignment` class defined in `trifusion.process.sequence`. This is an abstract class that contains the parsing methods (e.g. `_read_fasta`) for each supported input format. It is meant to be privately used within the context of the `AlignmentList` class, which provides the public main interface for handling alignment files.

There were three main challenges (goals) associated with the parsing procedure: (i) each alignment file had to be processed very quickly to allow tens of thousands of files to be parsed in a matter of seconds; (ii) the amount of sequence data that is stored in Random Access Memory (RAM) at each point in time had to be managed in order to parse very large alignments (> 1Gb) without storing the whole file in memory; and (iii) alignment data should be stored on disk for later retrieval with fast access. As I will describe below, these goals were achieved by leveraging the power of SQLite and a combination of lazy file generators and algorithms that minimizes memory usage and database operations.

4.4.1.1. SQLite database engine

In TriFusion, SQLite (https:www.sqlite.org) was used for data storage, manipulation and retrieval, via the `sqlite3` (https:docs.python.org/2/library/sqlite3.html) python module. In addition to the fact that SQLite was already used in the **Orthology** module, this SQL engine is fast, lightweight and provides several useful clauses for manipulating data.

After iterating over many table scheme strategies, the current version of TriFusion has a single main table and relies on 4 columns to keep record of some metadata besides taxon names and sequence strings. The definition of that table is as follows:

```
sql_cursor.execute(
    "CREATE TABLE [{}]("
    # Stores an unique integer identifier of a taxon
    "txId INT, "
    # Stores the taxon names
    "taxon TEXT, "
    # Stores the complete sequence string
    "seq TEXT, "
    # Stores a unique integer identifier of the alignment file
    "aln_idx INT".format(table_name, cols))

# Allows the creation of a index. By default, the index is
 ↪ created on the 'aln_idx' column to speed up the retrieval
 ↪ of data when querying by the alignment file.
sql_cursor.execute("CREATE INDEX {} ON [{}]({})".format(
    index[0], table_name, index[1]))
```

4.4.1.2. General parsing algorithm

All parsing methods, with the exception of *interleave* parsers (see section 4.4.1.4), follow the same general algorithm for inserting alignment data into an `sqlite` database. The key idea here is that files should be iterated using a lazy generator and sequence data should be inserted into the database as soon as possible. In fact, for all file formats, only a single taxon name and the corresponding sequence string are stored in memory at any given time. In this way, the peak memory footprint of parsing alignment files containing regular sized genes (0.5 – 10kb) is always under 10kb and even large concatenated sequences of $1e^7$ bp use only a few megabytes of RAM, regardless of the number of taxa and total size of the alignment file. Using the *phylip* parser as a template, the following simplified pseudocode is representative of the general parsing algorithm used in TriFusion:

```
size_list = []
for line in file_generator:
    # Get taxon and sequence strings from current line
    taxon, sequence = get_sequence_data(line)
    # Evaluate missing data symbol, if not yet set
    _eval_missing_symbol(sequence)
```

```
    # Insert data into the database
    _insert_data(taxon, sequence)
    # Append size of current sequence
    size_list.append(len(sequence))

# Check for sequence size consistency
_check_sizes(size_list)
# Add partition for current gene
_add_partition(**kwargs)
```

Alignment data is added to the database via a `INSERT INTO` statement in the `_insert_data` method:

```
try:
    # threading lock mechanism
    lock.acquire(True)

    self.cur.execute(
        "INSERT INTO alignment_data VALUES (?, ?, ?, ?)",
        (taxonID, taxon, seq, alnID))

finally:
    lock.release()
```

4.4.1.3. Parsing performance

As previously noted, the `AlignmentList` class is the public interface for loading alignment data into TriFusion's database. This class contains an `add_alignment_files` method that takes a list of alignment file paths as an argument and automatically parses and loads the data (alternatively, the list of file paths can be provided when instantiating the class). The performance of TriFusion's parsers can be assessed by profiling the `add_alignment_files` method. In the example below, a data set of 3093 alignments containing 141 taxa and a total of 212.6Mb of data is loaded and its parsing profiled.

```
from os import listdir
from os.path import join
from trifusion.process.sequence import AlignmentList

# Set the alignment file list
dataset_path = "path_to_3k_files"
dataset = [join(dataset_path, x) for x in
    ↪ listdir(dataset_path)]

# Load the alignments into TriFusion
aln_obj = AlignmentList(dataset, sql_path="db")
# Results of time and RAM profiling of the
    ↪ 'add_alignment_files' method
# Total time: 12.6952 s
# Peak RAM: 15.1 Mb
```

As can be seen from the profiling results, the parsing algorithm was able to process the data set in under 13s and with a peak RAM usage of 15Mb. On average, 34,000 taxon/sequence records can be parsed and loaded per second. Another useful profiling is the size of each `Alignment` instance containing several metadata information for each alignment file.

```
from pympler.asizeof import asizeof

average_size = sum(asizeof(x) for x in
    ↪ aln_obj.alignments.values()) / len(aln_obj.alignments)
print(average_size)
# 3055
```

On average, the `Alignment` objects contained in the `AlignmentList` instance were only roughly 3Kb each. In fact, this efficiency in RAM usage for each `Alignment` object scales particularly well with the size of the alignment file. For instance, if we load a single 1.1Gb file containing 40 taxa and perform the same profiling:

```
dataset_path = "path_to_1Gb_file"
dataset = [join(dataset_path, x) for x in
    ↪ listdir(dataset_path)]

```

```
# Load the alignments into TriFusion
aln_obj = AlignmentList(dataset, sql_path="db")
# Results of time and RAM profiling of the
  ↪ 'add_alignment_files' method
# Total time: 5.77852 s
# Peak RAM: 5.5Mb
aln = aln_obj.alignments.values()[0]
print(asizeof(aln))
# 3088
```

We can see that TriFusion was able to parse a 1.1Gb alignment file under 6 seconds and using only 5.5Mb of RAM. Despite the dramatic size increase relatively to the previous example, the size of the `Alignment` object of the new alignment remains just over 3Kb. In practice, this means that the alignment parsing algorithms implemented in TriFusion are able to scale very well with the increase of the alignment data set, both in terms of number of alignments and in the size of those alignments.

4.4.1.4. Interleave input formats

Interleave format variants pose a challenge to alignment parsing that is not trivial to solve if we aim to keep the same performance and efficiency of their *leave* parsers counter-parts. Due to the way sequence data is structured in these file formats, the complete sequence of any given taxon can only be gathered completely upon reaching the end of the file. This means that using the same algorithm of the *leave* formats would result in loading the entire alignment file into memory before the sequence data could be inserted into the database. Alternatively, we could force sequence data to be added into the database at every line, inserting or updating each database entry. However, updating an existing database row on each iteration is particularly costly in SQLite in comparison to the insertion of a new row. Since none of these two simpler modifications were satisfying in terms of performance and efficiency, I implemented a different algorithm when input alignments were detected as being in *interleave* format.

To prevent an *interleave* alignment file from being entirely loaded into memory while keeping the number of database insertions to a minimum, the file is parsed by iterating over a range that is determined by the number of taxa in the alignment. In each iteration, a new file object handle is created and we read over the entire file while retrieving only the sequence of a particular taxon. This ensures that, as in the *leave* parsers, only a single taxon and its corresponding sequence string are stored in memory at any given time. This also means that the alignment needs to be read N times, where N is the number of taxa. However, since the vast majority of lines are actually skipped in each iteration, the decrease in speed is marginal, while the gains in memory efficiency and performance in database operations are much larger. The following simplified pseudocode demonstrates the algorithm of the *interleave* parser.

```
for i in xrange(ntaxa):
    # Stores sequence strings
    sequence = []
    # Index identifier of the target taxon
    idx = 0
    # Sets when sequence should be gathered
    taxa_gather = True

    fh = open(self.path)

    for line in fh:
        # At blank lines, reset the idx and set taxa_gather to
        ↪ False.
        if not line.strip():
            idx = 0
            taxa_gather = False
        else:
            # Check if line is from target taxon
            if idx == i:
                # Remove the taxon from the gathering
                if taxa_gather:
                    sequence.append(
                        "".join(line.strip().split()[1:]))
                else:
                    sequence.append(
                        "".join(line.strip().split()))
            idx += 1

    seq = "".join(sequence)
```

```
29      taxa = self._taxa_idx.keys()[i]
30
31      # Insert data into database
32      self._insert_data(i, taxa, seq)
33      fh.close()
```

A noteworthy detail in the above code example is the storage of sequence fragments in a list, instead of performing successive concatenations with the previous string. Although this approach is not as intuitive as the concatenation one, it is well known that string concatenation is best done with `''.join(iter)` (which has $O_{(n)}$ complexity) than using the `+=` operator (which has $O_{(n**2)}$ complexity). Although more recent python versions have mitigated this issue, the `join` approach remains the best practice and can be orders of magnitude faster.

4.4.2 Data retrieval

With sequence data loaded in a single `sqlite` table, fast and efficient retrieval methods were required to perform all operations of TriFusion. In terms of performance, they key goal here is to minimize the number of queries made to the database for any particular task (ideally performing a single query). This is because the database will be able to perform better optimizations in single large queries and because there is always an overhead for communicating with it. To address these issues, we created two methods for the `AlignmentList` class named `iter_alignments` and `iter_columns`. These methods are lazy generators that yield, among other metadata, complete sequences and alignment columns, respectively. Overall, the optimizations implemented in these methods have a major impact on **all** of TriFusion's operations for the **Process** and **Statistics** modules, and allowed for a substantial and generalized speed up.

4.4.2.1. Complete sequence retrieval

The algorithm for the retrieval of complete sequence data for each taxon *and* alignment file was written in the `iter_alignments` method. A simplified version for demonstration purposes follows below:

```
def iter_alignments(self, table_name=None, include_txid=False):

    table_name = table_name if table_name else
     ↪  self.master_table

    try:

        lock.acquire(True)

        for txId, taxon, seq, aln_idx in self.cur.execute(
                "SELECT txId, taxon, seq, aln_idx "
                "FROM [{}] "
                "WHERE aln_idx NOT IN ({}) AND "
                "aln_idx IN ({})".format(
                    table_name,
                    ", ".join([str(x) for x in
                     ↪  self.shelved_idx]),
                    ", ".join([str(x) for x in
                     ↪  self.alignment_idx]))):
            if taxon not in self.shelved_taxa:
                if include_txid:
                    yield txId, taxon, seq, aln_idx
                else:
                    yield taxon, seq, aln_idx

    finally:
        lock.release()
```

This generator provides access to the complete set of taxon/sequence records contained in the `AlignmentList` object using a single query to the database. Since it is a generator, only the data for a single taxon/sequence is stored in memory at any given time during the iteration. For the purpose of fetching data, this method behaves like a regular iterable. For instance, the first sequence record in the database can be easily fetch with the following code:

```
from trifusion.process.sequence import AlignmentList

aln_obj = AlignmentList(ds, sql_db=sql_path)

# Get first taxon/sequence record
next(aln_obj.iter_alignments())
# (u'Schizophyllum_commune',
#  u'xxxxxxxxxxxxxxxxxxxxxma ... eavatrtkwftaxxx',
#  1)
```

As can be seen, the generator returns a tuple with the taxon name as the first element, the complete sequence as the second and the alignment file index as the third. In terms of performance, this query algorithm is also quite fast and provides only a minimal overhead. In the example above, we loaded the `ds` data set, which contains 3093 genes with 141 taxa and a total of ≈ 436k taxon/sequence records. If we profile the generator, we can see that it takes just over one second to iterate over the complete database and retrieve the sequence data from each taxon.

```
def query_db():
    for i in aln_obj.iter_alignments():
        pass

%lprun -f query_db query_db()
# Total time: 1.17637 s
```

In addition to its performance, this generator also takes into account the alignments and taxa in the active data sets. If the user modifies the active alignments and taxa, only those active elements will be yielded by the generator, therefore increasing the performance when processing custom active data sets.

```
# Set a single alignment as active
aln_obj.update_active_alignments(
    [os.path.join(d_path,
     ↪ "BasidioOnly2585_linsi_missingFilter_concPrep_big.fas")])
# Set a single taxon as active
aln_obj.update_taxa_names(["Postia_placenta"])

```

```
for record in aln_obj.iter_alignments():
    print(record)
# (u'Postia_placenta',
#  u'xxxxxxxxx ... xxxxxxxxxxx',
#  3083)
```

4.4.2.2. Alignment column retrieval

The efficient retrieval and iteration over alignment columns for each and every alignment file was considerably more challenging and required a more complex algorithm. The issue here is that sequence data that is contained in a single database must be grouped by the alignment file and then an array of every i^{th} column of a particular alignment must be yielded. Using an algorithm similar to the one defined for the `iter_alignments` generator would imply the storage of the entire data set in memory, so an entirely new database query was devised. The following simplified code provides a demonstration of the alignment column retrieval algorithm:

```
def iter_columns(self, table_name=None, aln_idx=None,
                 include_taxa=False, group_by=None):

    query = "SELECT " \
            "{tx} " \
            "GROUP_CONCAT(substr(seq, {pos}, 100000))"\
            "{idx} " \
            "FROM [{tb}] " \
            "WHERE {cond} " \
            "AND {cond_tx} " \
            "GROUP BY {idx}"

    shelved = ", ".join([str(x) for x in self.shelved_idx])
    cond = "aln_idx NOT in ({})".format(shelved)

    cond_tx = "taxon NOT IN ({})".format(", ".join(
        ["'{}'".format(x) for x in self.shelved_taxa]))

    tx_query = "GROUP_CONCAT(taxon),"

    for p in xrange(0, self.size, 100000):
        for r in ((z, x.split(","), y) for z, x, y in
```

```
                    self.cur.execute(
                        query.format(
                            pos=p,
                            tb=table_name,
                            cond=cond,
                            cond_tx=cond_tx,
                            tx=tx_query,
                            idx=group_idx))):
            for col in itertools.izip(*r[1]):
                yield r[0].split(","), col, r[2]
```

The key features of this algorithm are the use the `izip` iterator from the `itertools` module and the use of the `subst` and `GROUP_CONCAT` functions of `sqlite`. The `izip` iterator aggregates the elements from each of the iterables at every i^{th} position, therefore yielding the alignment column, but it only evaluates the results upon request. This is important because it will only hold in memory the data from a single alignment column at any given time. However, the `izip` iterator needs a list of iterables to work with. Instead of providing the entire data set as input, which would mean loading it in memory, a combination of the `substr` and `GROUP_CONCAT` functions in `sqlite` were used to provide an alignment block of a specific size. I have currently set this block size to 100Kbp, as subsequent testing demonstrated that it provided a good trade-off between performance and memory usage. In short, this generator successively slices blocks of 100Kbp from the entire data set and feeds each block to the `izip` iterator, which yields an alignment column.

Using this algorithm, alignment columns can be easily iterated over like in the `iter_alignments` generator:

```
from trifusion.process.sequence import AlignmentList

aln_obj = AlignmentList(ds, sql_db=sql_path)
next(aln_obj.iter_columns())
# ((u'x', u't', u'm', ...   u'm',  u'm',  u'x'),
    1)
```

If the assignment of each element in the alignment column to the corresponding taxon name is important, we can use the `include_taxa` option:

```
next(aln_obj.iter_columns(include_taxa=True))
# ([u'Schizophyllum_commune',  u'Trametes_versicolor', ... ,
 ↪  u'2Yarrowia_lipolytica2'],
#  (u'x', u't', ... , u'x'),
#  1)
```

In terms of performance, iterating over the columns of a data set with 3093 genes, 141 taxa and a total of ≈ 213k alignment columns takes about 9 seconds to complete.

```
def query_db():
    for i in aln_obj.iter_columns():
        pass

%lprun -f query_db query_db()
# Total time: 9.50347 s
```

4.4.3 Main operations

4.4.3.1. Concatenation

An algorithm to perform the concatenation of multiple alignment files in a single one is conceptually easy to devise but challenging to optimize. In addition to the fact that absent taxa in particular alignments need to be filled with missing data, sequence data from all alignments needs to be merged together without being stored in memory.

To address this issue, I devised a two-step concatenation algorithm in the `concatenate` method of the `AlignmentList` class. First, a temporary `sqlite` table will be populated with the sequence records of all taxon/sequence *plus* the missing data sequences of taxa that are absent in any given alignment file. It is important that this temporary table has an index created for the `txId` column,

which will store the identifier of each taxon in the data set for the second part of the algorithm.

```
# Create temporary table and create an index
temp_table = ".concatenation"
self._create_table(temp_table, index=["concidex", "txId"])

prev_idx = ""
for taxon, seq, aln_idx in self.iter_alignments():

    # This happens when the alignment changes during the
     ↪ iteration.
    if aln_idx != prev_idx:

        # If prev_idx is already defined, it means this is the
         ↪ second
        # alignment during the iteration. In that case, fill
         ↪ the table
        # with missing data from the previous alignment.
        if prev_idx:
            fill_missing_taxa(aln_obj, conc_cur)

        prev_idx = aln_idx
        aln_obj = self.alignment_idx[prev_idx]

    # Add data to temporary table
    conc_cur.execute(
        "INSERT INTO [{}] VALUES (?, ?, ?, ?)".format(
            temp_table), (taxa_idx[taxon], taxon, seq, 1)
    )
```

As can be seen, whenever the iteration changes the alignment file (that is, when `aln_idx != prev_idx`), the `fill_missing_taxa` function is called to fill the database with missing sequences for all taxa that were absent from the previous alignment. Therefore, this new table already includes the complete set of sequences that will appear in the final concatenation file, including missing data sequences.

```
# Original table
# (1, "txA", "AAA", 1)
# (2, "txB", "AAA", 1)
# (1, "txA", "AAA", 2)

# (1, "txA", "AAA", 3)
# (2, "txB", "AAA", 3)
#
```

```
# Temporary table
# (1, "txA", "AAA", 1)
# (2, "txB", "AAA", 1)
# (1, "txA", "AAA", 2)

# (2, "txB", "NNN", 2)
# (1, "txA", "AAA", 3)
# (2, "txB", "AAA", 3)
```

We now only need to aggregate sequences with the same taxon ID across all alignment files, which is the second part of the algorithm:

```
for p, (idx, tx, seq, aln_idx) in enumerate(conc_cur.execute(
        "SELECT txId, taxon, GROUP_CONCAT(seq, ''), aln_idx "
        "FROM [{}] "
        "GROUP BY txId".format(temp_table))):

    self.cur.execute("INSERT INTO [{}] VALUES (?, ?, ?, ?)".format(
        table_out), (idx, tx, seq, aln_idx))
```

In the code above, we leverage the combination of the `GROUP_CONCAT` function with the `GROUP BY` clause of `sqlite` to perform the aggregation of sequence data by each taxon. The definition of an index on the column targeted by `GROUP BY` and performing the string concatenation using `sqlite` functions was crucial for the performance and efficiency of this operation. In fact, performing this string aggregation in python using either the `join` or `+=` methods was not only substantially slower but it also increased the memory consumption to approximately the size of a single concatenated sequence.

Profiling the concatenation method, we can see that it only takes around 9 seconds to concatenate a data set of 3093 genes and 141 taxa (213Mb) with a peak memory usage of 22Mb.

```
%lprun -f aln_obj.concatenate aln_obj.concatenate()
# Total time: 8.6646 s
%mprun  -f aln_obj.concatenate aln_obj.concatenate()
# Peak memory: 22 Mb
```

4.4.3.2. Reverse concatenation

The reverse concatenation of a single alignment file entails the successive slicing of alignment blocks according to a user defined partition scheme. This was one of the most technically challenging operations in TriFusion for several reasons. First, it relies heavily on string operations, which are quite computationally expensive. Second, since TriFusion supports the reverse concatenation of codon partitions, some of the string slices need to be interspersed (e.g. at every 3^{rd} character). Finally, an efficient strategy for querying the database with a low number of operations was not obvious.

As in the concatenation algorithm, I devised a two-step algorithm to handle the reverse concatenation operation implemented in the `reverse_concatenate` method of the `AlignmentList` class. In the first step, a temporary `sqlite` table with an index on the alignment file identifier is created. Then, TriFusion iterates only once over each sequence in the database and keeps a single sequence in memory at any given time. For each partition range, the current sequence is sliced and inserted into the database with a new alignment file identifier that matches the partition index. A simplified version of the algorithm follows below:

```
temp_table = ".reversedata"
self._create_table(temp_table, index=("revindex", "aln_idx"))

prev_idx = ""
for p, (taxon, seq, aln_idx) in enumerate(
        self.iter_alignments(table_in)):

    if prev_idx != aln_idx:

```

```
    # Index for partitions
    part_idx = 1

    for name, part_range in self.partitions:

        name = part_map[name]

        if part_range[1]:

            # Do codon partition slice

            temp_cur.execute(
                    "INSERT INTO [{}] VALUES (?, ?, ?, ?)".format(
                        temp_table), (p, taxon, part_seq,
                         ↪  part_idx))

        else:

            # Do single partition slice

            temp_cur.execute(
                "INSERT INTO [{}] VALUES (?, ?, ?, ?)".format(
                        temp_table), (p, taxon, part_seq,
                         ↪  part_idx))
```

The new temporary table will hold the taxon/sequence records for the individual partitions, albeit unsorted. The second part of the algorithm involves the re-ordering of the table, the creation of the new `Alignment` objects and modification of the `AlignmentList` attributes:

```
self._create_table(table_out, index=("finalrevindex",
 ↪  "aln_idx"))
self.cur.execute(
    "INSERT INTO [{}] (txId, taxon, seq, aln_idx) "
    "SELECT txId, taxon, seq, aln_idx "
    "FROM [{}] "
    "ORDER BY aln_idx".format(table_out, temp_table))

part_idx = 1
for name, part_range in self.partitions:

    name = part_map[name]

```

```
    if part_range[1]:

        for i in range(3):

            # Create codon Alignment
            # Modify AlignmentList attributes

    else:

        # Create single Alignment
        # Modify AlignmentList attributes
```

Note in the above code that TriFusion creates the final table and inserts the data for query as it performs the `ORDER BY` clause by the `aln_idx` column (which stores the partition identifier set in the first part of the algorithm). This ensures that the reverse concatenation algorithm is performed with only two database calls while storing a single sequence record in memory at any given time. However, the modifications that the reverse concatenation operations make on the `AlignmentList` object also required a substantial modification of its attributes (see the source for the full method).

Considering a concatenated file of 48 taxa and a partition scheme of 3093 partitions, the `reverse_concatenate` method can reverse this alignment into 3093 individual alignments in 13 seconds and with a peak memory usage of 22Mb.

```
from trifusion.process.sequence import AlignmentList

aln_obj = AlignmentList(ds, sql_db=sql_path)
aln_obj.partitions.read_from_file(partition_path)

%lprun -f aln_obj.reverse_concatenate
 ↪  aln_obj.reverse_concatenate()
# Total time: 13.0549 s
%mprun -f aln_obj.reverse_concatenate
 ↪  aln_obj.reverse_concatenate()
# Peak memory: 22 Mb
```

4.4.4 Secondary operations

4.4.4.1. Collapse

In the collapse operation, identical sequences across two or more taxa need to be collapsed into a single sequence. The challenged posed by this operation is that each sequence has to be compared with each other, which could potentially lead to the entire data set to be loaded into memory.

The algorithm I implemented in the `collapse` method of the `AlignmentList` class bypasses the need to compare full sequence strings by hashing each sequencing and comparing only the hashes. Not only is hash comparison much faster than comparing two sequence strings that could be megabases of length, but is also ensures that only a single sequence string is loaded in memory before creating its hash. Whenever a new hash is read from the database, the corresponding sequence is stored in a new `sqlite` table and the hash stored in a dictionary for future comparisons. When a new sequence's hash already exists in the hash list, we only take note that the current taxon has the same sequence.

```
temp_table = ".collapsed"
self._create_table(temp_table)

prev_idx = ""
for taxon, seq, aln_idx in self.iter_alignments(table_in):

    if aln_idx != prev_idx:

        hash_list = {}
        hap_dic = {}
        hap_counter = 0

        prev_idx = aln_idx

    cur_hash = hash(seq)

    if cur_hash not in hash_list:

```

```
        haplotype = "{}_{}".format(haplotype_name, hap_counter +
        ↪ 1)
        hash_list[cur_hash] = haplotype
        hap_dic[haplotype] = [taxon]

        temp_cur.execute(
            "INSERT INTO [.collapsed] VALUES (?, ?, ?, ?)",
            (hap_counter, unicode(haplotype), seq, aln_idx))

        hap_counter += 1

    else:

        hap_dic[hash_list[cur_hash]].append(taxon)
```

As the above code demonstrates, the use of sequence hashes greatly simplifies comparisons across sequences in the same alignment file. If we collapse a single alignment with 376 taxa and ≈ 1.1Gb, the `collapse` method is able to perform the operation in under 5 seconds and using less than 28Mb of RAM (note that each sequence in the alignment has ≈ 1.5Mb).

```
from trifusion.process.sequence import AlignmentList

aln_obj = AlignmentList(ds, sql_db=sql_path)
%lprun -f aln_obj.collapse aln_obj.collapse()
# Total time: 4.8369 s
%mprun -f aln_obj.collapse aln_obj.collapse()
# Peak memory: 27.1 Mb
```

4.4.4.2. Filter by missing data

TriFusion allows the filtering of alignment columns by the amount of missing data present in them. One of the issues in this operation is that it implies the iteration of data by alignment column instead of by each sequence, which means a substantially larger number of iterations. Moreover, sequence modifications in this operation needed to be performed at the base/residue level, which means a larger number of operations per sequence and taxa. Earlier attempts of creating temporary empty

strings or lists for each taxon and then populating those objects as the alignment columns were being iterated resulted in poor performance for large alignments (>100Kb). Both the `join` and `+=` operators had scaling problems with the increase in sequence size.

To solve this issue, we implemented an algorithm in the `_filter_columns` method of the `AlignmentList` class that creates a binary array with the length of the sequence and scores `1` for "good" columns. Being a binary array, it is very cheap to build and only one needs to be created for each alignment. Then, we used the `compress` function from the `itertools` module to filter elements from an iterable (the sequence string) and return only those with a `True` value. Delegating the heavy-lifting of string compression to an optimized function yielded a performance improvement many times higher than the previous approach. More importantly, this approach is able to scale particularly well even with the increase of sequence size.

Considering a data set of 3093 alignments and 48 taxa, the `filter_missing_data` method, which is the public interface for missing data filtering and calls `_filter_columns`, was able to filter sequences in roughly 62 seconds and using only 10 Kb of memory.

```
from trifusion.process.sequence import AlignmentList

aln_obj = AlignmentList(ds, sql_db=sql_path)
%lprun -f aln_obj.filter_missing_data
 ↪  aln_obj.filter_missing_data(50, 75)
# Total time: 62.2172 s
%lprun -f aln_obj.filter_missing_data
 ↪  aln_obj.filter_missing_data(50, 75)
# Peak memory: 10 Kb
```

Despite the new optimized algorithm, filtering alignments by their columns remains an intensive task that is bound to take longer than other operations that perform fewer iterations over complete sequence data.

4.4.4.3. Filter by codon position

This operation shared the same issues of the missing data filter from the previous section, but required a slightly different approach to compress the sequence string. In this operation we needed to filter alignment columns with a known periodicity (every 3^{rd} character). Therefore, we defined a binary generator to create the binary array used for `compress`.

```
def index(length, pos):
    """
    index generator
    """
    for _ in range(0, length, 3):
        for j in pos:
            if j:
                yield 1
            else:
                yield 0

prev_idx = ""
for txId, taxon, seq, aln_idx in self.iter_alignments(
        table_in, include_txid=True):

    if aln_idx != prev_idx:

        if prev_idx:
            aln_obj.locus_length = len(final_seq)
            self.set_partition_from_alignment(aln_obj)

        prev_idx = aln_idx

    final_seq = "".join(list(itertools.compress(
        seq, index(aln_obj.locus_length, position_list))))

    temp_cur.execute(
        "INSERT INTO [{}] VALUES (?, ?, ?, ?)".format(temp_table),
        (txId, taxon, final_seq, aln_idx))
```

The `index` generator ensures that the required binary array is not even stored in memory during this operation and greatly simplifies the compress procedure. As in previous operations, only a sequence record is stored in memory at every given

time point. Removing the 3^{rd} codon position from each alignments on a data set of 3093 genes and 48 taxa takes:

```
from trifusion.process.sequence import AlignmentList

aln_obj = AlignmentList(ds, sql_db=sql_path)
```

4.4.5 Writers

Having stored sequence data in a single `sqlite` table with a specific schema and ensuring that all operations that modify that database maintain that schema greatly simplifies the work of the writer methods. There are methods for each supported output format in the `AlignmentList` class (e.g. `_write_fasta`) and most follow the same straightforward approach. Taking the template of the phylip output format from the `_write_phylip` method:

```
for taxon, seq, aln_idx in self.iter_alignments(table_name):

    if aln_idx != prev_file:
        prev_file = aln_idx
        fh, of = self._setup_newfile(
            fh, aln_idx, output_dir, suffix, output_file, ns)

        if not fh:
            continue

       aln_obj = self.alignment_idx[aln_idx]

       self._write_phylip_partitions(aln_obj, partition_file,
                                     of, model_phylip)

        fh.write("{} {}\n".format(
            len(aln_obj.taxa_idx) - len(aln_obj.shelved_taxa),
            aln_obj.locus_length))

    if not fh:
        continue

    fh.write("{} {}\n".format(
```

```
24          taxon[:cut_space_phy].ljust(tx_space_phy),
25          seq))
```

As can be seen from the code above, writing all output files requires only one database query. We have defined methods that handle the setup of each output file, optional partitions files and, therefore, the process of writing the sequence data into each file is simple, fast and efficient.

However, there are two scenarios that required more complex algorithms and modifications to the database in order to produce output files with similar speed and efficiency: **Interleave formats** and **partitioned formats**, including data sets that mix nucleotide and protein data.

4.4.5.1. Interleave formats

The challenge with the *interleave* output format is similar to the corresponding input format. By default, sequence data is stored in *leave* format in the database but the output files must have sequences truncated to 90bp and each taxon must appear interspersed throughout the entire file. The most simple approach to this problem would be to iterate the database over N times, where $N = \frac{alignment\ size}{90}$, and in each iteration write the truncated sequence for every taxon. The obvious issues with this approach are the large number of database queries and the strong reliance on string operations.

In TriFusion, we address this challenge by creating a temporary `sqlite` table, where the truncated sequences are inserted sequentially for each taxon at a time. Then we use `sqlite` operations to re-arrange the table in such a way that it only takes a single query to write all output files. This algorithm is implemented in the `_get_interleave_data` method of the `AlignmentList` class:

```
self.cur.execute("CREATE TABLE [.interleavedata]("
                 "taxon TEXT,"
                 "seq TEXT,"
                 "slice INT,"
                 "aln_idx INT)")
self.cur.execute("CREATE INDEX interindex ON "
                "[.interleavedata](aln_idx, slice)")

temp_cur = self.con.cursor()

prev_file = ""
for taxon, seq, aln_idx in self.iter_alignments(
        table_name=table_name):

    for i in xrange(90, aln_obj.locus_length, 90):

        temp_cur.execute("INSERT INTO [.interleavedata] VALUES "
                         " (?, ?, ?, ?)", (taxon,
                          ↪  seq[counter:i], i,
                                           aln_idx))
        counter = i

    try:
        if aln_obj.locus_length % 90:
            i += 1
            temp_cur.execute(
                "INSERT INTO [.interleavedata] VALUES "
                " (?, ?, ?, ?)", (taxon, seq[counter:],
                                  i, aln_idx))
    except UnboundLocalError:
        temp_cur.execute("INSERT INTO [.interleavedata] VALUES "
                         " (?, ?, ?, ?)", (taxon, seq, 0,
                          ↪  aln_idx))
```

A key aspect of the code above is the creation of an index on the `aln_idx` and `slice` columns for the temporary table. These are the columns that will be later used in the query to quickly and efficiently re-arrange the table into *interleave* format. Using the template of the phylip format, the single query for writing sequence data becomes simply:

```
for taxon, seq, p, aln_idx in self.cur.execute(
    "SELECT taxon, seq, slice, aln_idx from [.interleavedata] "
    "ORDER BY aln_idx, slice"):
```

Profiling the `write_to_file` method when setting the output format to *phylip* and specifying the *interleave* format for a data set of 3093 alignments and 48 taxa reveals that TriFusion performs this operation in 17 seconds and under 10kb of memory.

```
from trifusion.process.sequence import AlignmentList

aln_obj = AlignmentList(ds, sql_db=sql_path)

%lprun -f aln_obj.write_to_file
    aln_obj.write_to_file(["phylip"],
    output_dir="/home/diogo/Diogo/teste/", interleave=True)
# Total time: 17.2067 s
%mprun -f aln_obj.write_to_file
    aln_obj.write_to_file(["phylip"],
    output_dir="/home/diogo/Diogo/teste/", interleave=True)
# Peak memory: 10 Kb
```

4.4.5.2. Partitioned data sets

Some output formats (e.g. MCMCTree) require the output file to be formatted according to a partition scheme. Moreover, a mix of nucleotide and protein data may be loaded in TriFusion and it is necessary to merge the sequence types together for several output formats and downstream analyses. The first challenge is very similar to the *interleave* output format algorithm with the difference that sequences are sliced according to partition ranges instead of at 90bp. The second challenge adds a new layer of complexity (the sequence type of the partitions) that must be taken into account when re-arranging the database table.

As in the *interleave* output format, I devised an algorithm in the `_get_partition_data` method of the `AlignmentList` class that creates a new `sqlite` table and re-arranges its data. To handle the different sequence types, additional columns were included into the new table and an index with an additional column was created. Then, data was inserted into the database sequentially for each taxon/sequence record according to the defined partitions:

```
partition_table = ".partitiondata"
self.cur.execute("CREATE TABLE [{}]("
                 "txId INT,"
                 "taxon TEXT,"
                 "seq TEXT,"
                 "part_name TEXT,"
                 "part INT, "
                 "aln_idx INT, "
                 "part_type INT)".format(partition_table))

self.cur.execute("CREATE INDEX partindex ON "
                 "[{}](part_type, part, aln_idx)".format(
    partition_table))

prev_tx = ""
for txId, taxon, seq, aln_idx in self.iter_alignments(
        table_name, include_txid=True):

    for name, part_range in part_lst.items():

        # Get partition sequence and metadata

        temp_cur.execute(
            "INSERT INTO [{}] "
            "VALUES (?, ?, ?, ?, ?, ?, ?)".format(
                partition_table),
            (txId, taxon, part_seq, nm, part_idx, aln_idx,
             ↪  seq_type))
```

Note that with the index defined for the new table, we take into account the sequence type (`part_type`), the partition index (`part`) and the alignment file index (`aln_idx`). This allows the table to be quickly and efficiently sorted into the required format for partitioned output formats and already has potential nucleotide and protein sequences merged together. Therefore, if querying for partitioned output formats such as MCMCTree, querying this table becomes:

```
for taxon, seq, pname, pidx, aln_idx in self.cur.execute(
        "SELECT taxon, seq, part_name, part, aln_idx "
        "FROM [.partitiondata] "
        "ORDER BY aln_idx, part"):
```

On the other hand, re-arranging the table for regular output formats that contain a mix of nucleotide and protein sequences can be quickly performed with:

```
temp_cur.execute(
    "INSERT INTO [{}] "
    "(txId, taxon, seq, aln_idx) "
    "SELECT txId, taxon, GROUP_CONCAT(seq, ''), aln_idx "
    "FROM (SELECT txId, taxon, seq, part, aln_idx FROM [{}] "
    "ORDER BY part_type, part)"
    "GROUP BY txId".format(table_name, partition_table))
```

4.5 Continuous integration, unit testing and code quality

As the code base of TriFusion grew (currently sitting at more than 100k lines of code), it became critical to setup a robust, complete (as much as possible) and independent suite of tests that ensured the reliability of the distributed software throughout development. In fact, given the complexity and size of the project, this suite of tests are essential for an agile and sustainable development process across multiple machines and operating systems. This would allow substantial changes to be undertaken in the code base, such as adding a new feature or fixing a complex bug, and immediately know if the remaining code is working as expected. If not, the failed tests provide the necessary information about what went wrong and where, so that the problem can be fixed before merging the changes into the main branch.

One strategy to achieve this is by using a Continuous Integration (CI) approach, which is the automated process of building and testing the code every time a commit is pushed to the version control system. With TriFusion, we leveraged the combination of the modern version control system, GitHub (https://github.com/), with a hosted, distributed and CI service called Travis CI (https://travis-ci.com/). A comprehensive suite of unit tests was written using the *de facto* testing framework for python, `unittest` (https://docs.python.org/2/library/unittest.html), and integrated within the CI service. In addition, we also integrated other services

that measure the coverage of the code that is being tested with CodeCov (https://codecov.io/) and the overall code quality of the repository with Codacy (https://www.codacy.com/).

4.5.1 Continuous integration

TriFusion's source code is currently hosted at GitHub (https://github.com/ODiogoSilva/TriFusion), which has pre-built integration and service hooks with the Travis CI service. The configuration of Travis is done via a YAML file at the root of the repository named `.travis.yml`.

```
language: python

python:
    - "2.7"

before_install:
    - sudo apt-get update -q
    - sudo apt-get install libblas-dev liblapack-dev gfortran

install:
    - pip install scipy numpy matplotlib pandas
      ↪ python-coveralls coverage nose
    - python setup.py install

script:
    - export PYTHONIOENCODING=UTF-8
    - nosetests --with-coverage

after_success:
  - bash <(curl -s https://codecov.io/bash)  -g
    ↪ **/progressbar/*
```

This configuration file contains all the necessary instructions and installation steps that are required to build and test TriFusion's source on an independent machine. The fact that this is done on an independent machine that builds the whole environment from scratch is of great importance because it may uncover environmental problems that are difficult to identify in the development box. For instance, the

application may install and run successfully on the development environment due to the existence of a given library that is not specified as a dependency because the developer is unaware of it. This CI service prevents the occurrence of these situations by notifying the developer of any errors during the build process.

Once setup, this building process is triggered at each commit to version control and is automatically executed on Travis CI servers. Its most important feature is the continuous reporting and notifications (Figure 4.2), particularly when something stops working or is not working as expected.

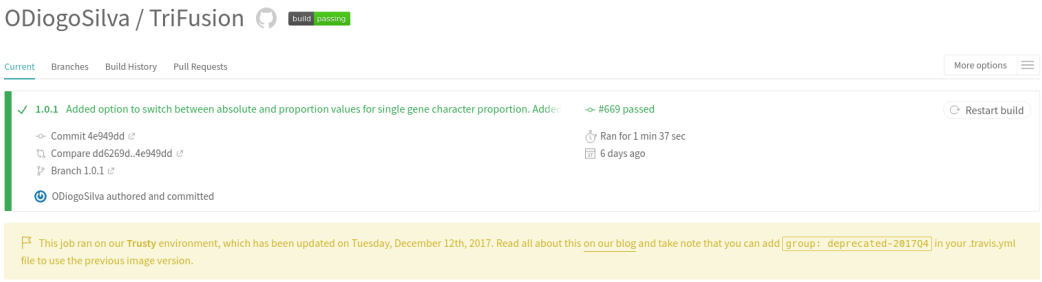

Figure 4.2. Example of the build and testing status of Travis CI for TriFusion.

Another important feature is the possibility of integrating other services during the building process. For instance, the suite of unit tests of the TriFusion repository (section 4.5.2) can be automatically executed after a successful build by specifying the appropriate command in `.travis.yaml` (See line 15). Once the tests have been completed, a code coverage analysis is automatically performed (See line 19)

4.5.2 Unit/integration tests and code quality

Unit testing is the practice of testing individual units of code to verify if they are working as expected. These units are short and (ideally) independent code segments that can be viewed as the smallest testable parts of an application. In each unit, it is possible to provide a given set of inputs to a certain function and then assert whether it is returning the proper values and/or behaving as expected. Integration tests is when multiple modules/functions are tested in combination to ensure that each module is correctly integrated. A widely regarded advantage of writing these tests for an application is that it motivates the developer to write code that is more

modular and easier to test instead of large and complex pieces of code that are harder to test and maintain. Naturally, this comes at the expense of investing time on writing a suite of tests during the development, but it also provides a wealthy number of advantages, most notably the fact that it could prevent future changes from breaking functionality. Moreover, once written, the tests can be automatically executed with minimal effort. The output of a test is simply a console output with the result of the test assertion, which can be a `pass` or `fail` condition. The `pass` condition is uninteresting but a good indication of the function's health. The `fail` condition alerts the developer of an error or unexpected behavior on that particular function and pinpoints the source of the problem when the tests are well designed. Ultimately, when a comprehensive suite of tests that covers the majority of the code base has been written, passing all tests provides great confidence that the entire application is working correctly.

In TriFusion, a suite of unit and integration tests was written using the `unittest` framework from the Python standard library (https://docs.python.org/2/library/unittest.html). Among its many features, `unittest` supports test automation, the setup and shutdown of groups of test cases, the aggregation of tests into collections and the ability to collect information about testing coverage. The development of this test suite is an ongoing process that is updated every time new functionalities are added, but 228 tests have been currently written so far (See https://github.com/ODiogoSilva/TriFusion/tree/master/trifusion/tests). To demonstrate how tests have been implemented in TriFusion and how they can be useful, we could consider a test case for the alignment filtering methods. First, we create a new `TesteCase` specifying how each individual test should be setup and terminated.

```
class AlignmentMissingFiltersTest(unittest.TestCase):

    def setUp(self):

        if not os.path.exists(temp_dir):
            os.makedirs(temp_dir)

        self.aln_obj = AlignmentList([], sql_db=sql_db)
```

```
9
10      def tearDown(self):
11
12          self.aln_obj.clear_alignments()
13          self.aln_obj.con.close()
14          shutil.rmtree(temp_dir)
```

In the example above, we specify what is done before the actual test in the `setUp` method, which, in this case, creates a temporary directory and an empty `AlignmentList` object. At the end of each test, the `tearDown` method is called, clearing the `AlignmentList` object, closing the connection to the database and removing the temporary directory.

Now consider TriFusion's functionality of filtering alignments according to missing data content in each alignment column. First, we could test the scenario where the filtering function is called with the maximum possible value for the missing data and gap filters, which actually means that no columns should be filtered. The expected behaviour in this case is that the final alignment should be identical to the input alignments.

```
1  def test_no_filters(self):
2
3      self.aln_obj.add_alignment_files(
4          ["trifusion/tests/data/missing_data.phy",
5           "trifusion/tests/data/missing_data2.phy"]
6      )
7
8      self.aln_obj.filter_missing_data(100, 100)
9
10     s = []
11     for aln in self.aln_obj:
12         s.append(aln.locus_length)
13
14     self.assertEqual(s, [50, 50])
```

As can be seen in the example above, two alignment files are provided as input and the filter method (`filter_missing_data`) is called with the maximum filter values.

After the filtering is performed, we assert whether the final alignments contain the same length as the original ones (50 base pairs). We could also test the opposite scenario, where the filter values are at their minimum, which means that the final alignment should not contain any missing data or gaps.

```
def test_no_missing(self):

    self.aln_obj.add_alignment_files(
        ["trifusion/tests/data/missing_data.phy",
         "trifusion/tests/data/missing_data2.phy"]
    )
    self.aln_obj.filter_missing_data(0, 0)

    s = []
    for aln in self.aln_obj:
        s.append(aln.locus_length)

    self.assertEqual(s, [0, 19])
```

In this case, we know that the first alignment would become empty and the second would retain only 19 alignment columns, so we use these values to test our assertion. Having these tests setup and executed automatically will ensure that future modifications in the code that impact the alignment filtering method can be verified and breaking changes will fail one or more of these tests.

Another issue that arose when developing TriFusion's was the necessity of comparing the attributes between two Class object instances. As previously described (See section 4.4.1), the majority of TriFusion's code is Class based and relies on Object Oriented Programming to make objects interact with one another. As such, it was useful to compare the state of the attributes between two instances in some cases. For instance, when an `AlignmentList` is cleared we need to make sure that all attributes were reset to the original state. We achieved this by extending the default assertion test of the `unittest` framework.

```
def compare_inst(inst1, inst2, blacklist=None):

    d1 = dict((x, y) for x, y in inst1.__dict__.items() if x
     ↪  not in blacklist)
    d2 = dict((x, y) for x, y in inst2.__dict__.items() if x
     ↪  not in blacklist)

    try:
        return d1 == d2
    except ValueError:
        return any(d1) == any(d2)
```

Then, this function could be provided to any test assertion, comparing each attribute between two arbitrary instances except for the ones provided via the `blacklist` argument.

```
def test_clear_alns(self):

    self.aln_obj.clear_alignments()
    aln = AlignmentList([], sql_db=sql_db)

    self.assertTrue(compare_inst(self.aln_obj, aln,
         ["log_progression",
          "locus_length",
          "partitions",
          "cur",
          "con"]))
```

With this test, we can always know if the `clear_alignments` method is resetting the `AlignmentList` object to the original state (the blacklisted attributes are actually references to other Class objects and need to be compared elsewhere). This is quite useful when a new attribute is added to the Class but we forget to reset it when clearing the instance.

An apparent limitation of the `unittest` framework is that each test is limited to a single assertion. While this is the correct and expected conceptual behaviour for unit tests, there are some cases where making multiple assertions in a single test is more useful than writing a multiple nested test. For example, considering

the case where we need to test whether partitions are being correctly removed from the `AlignmentList` object, we would first need to test whether the partition (or partitions) was correctly removed from the instance's attributes, which would take up one assertion. However, in addition to this assertion, we also wanted to check if the ranges of the remaining partitions were continuous, since the removal of intermediary partitions would affect the ranges of all subsequent ones. To work around this limitation, we wrote a custom `TestCase` subclass that extends the normal behaviour of assertion tests to allow for multiple assertions and report which ones failed.

```
class ExpectingTestCase(unittest.TestCase):
    def run(self, result=None):
        self._result = result
        self._num_expectations = 0
        super(ExpectingTestCase, self).run(result)

    def _fail(self, failure):
        try:
            raise failure
        except failure.__class__:
            self._result.addFailure(self, sys.exc_info())

    def expect_true(self, a, msg):
        if not a:
            self._fail(self.failureException(msg))
        self._num_expectations += 1

    def expect_equal(self, a, b, msg=''):
        if a != b:
            msg = '({}) Expected {} to equal {}. '.format(
                self._num_expectations, a, b) + msg
            self._fail(self.failureException(msg))
        self._num_expectations += 1
```

The custom `ExpectingTestCase` Class inherits from `TestCase`. It then extends the behaviour of the `run` method, overwrites the `_fail` method and provides two new assertion methods: `expect_true` and `expect_equal`.

We can now create test cases using this Class and perform multiple assertions in its individual tests.

```
class PartitonsTest(ExpectingTestCase):

    def setUp(self):

        if not os.path.exists(temp_dir):
            os.makedirs(temp_dir)

        self.aln_obj = AlignmentList(dna_data_fas,
         ↪  sql_db=sql_db)
        self.aln_obj.partitions.reset(
            cur=self.aln_obj.cur,
            keep_alignments_range=True
        )

    def tearDown(self):

        self.aln_obj.clear_alignments()
        self.aln_obj.con.close()
        shutil.rmtree(temp_dir)

    def test_remove_partition_from_name(self):

        self.aln_obj.partitions.read_from_file(
            concatenated_small_parNex[0],
            no_aln_check=True
        )
        self.aln_obj.partitions.remove_partition(
            "BaseConc3.fas"
        )

        key_data = [
            list(self.aln_obj.partitions.partitions.keys()),
            list(self.aln_obj.partitions.
                    partitions_alignments.keys()),
            list(self.aln_obj.partitions.models.keys())
        ]

        self.expect_equal(key_data,
                          [["BaseConc1.fas", "BaseConc2.fas",
                           "BaseConc4.fas",
                           "BaseConc5.fas", "BaseConc6.fas",
                           "BaseConc7.fas"]] * 3)

        cont = True
        prev = 0
        for r in self.aln_obj.partitions.partitions.values():
            if r[0][0] == prev:
                prev = r[0][1] + 1
```

```
48                 else:
49                     cont = False
50 
51             self.expect_equal(cont, True)
```

As seen above, the `test_remove_partition_from_name` test first verifies if the `partitions`, `partitions_alignments` and `models` attributes were correctly updated after removing a given partition. If that assertion passes, it proceeds to the next test where we test if the new partition ranges are correctly continuous. More generally, with this custom Class we can always use the default assertions or any number of multiple assertion in any individual test.

Overall, the addition of a unit test suite to the TriFusion package was a time consuming but worthwhile endeavour. The current suite now covers ~ 87% of the testable code (https://codecov.io/gh/ODiogoSilva/TriFusion), which is a substantial proportion for a project of this size. However, one limitation remains, which is the testing of GUI components of the `Kivy` framework, particularly due to the dependency of `OpenGL`. Unfortunately, `Kivy` lacks a testing framework and a new one would have to be written from scratch, possibly using the Model-view-controller (MVC) architectural pattern. We found that this endeavour would be difficult to pursue in the proposed time frame for the development of TriFusion, so instead we focused on expanding the test suite of the core and backend functionalities of TriFusion, and on good code quality and practices. Indeed, the code quality of TriFusion is tracked as part of the CI process, which grants the maximum quality certification to TriFusion (Figure 4.3; https://www.codacy.com/app/o.diogosilva/TriFusion/dashboard).

4.6 Benchmarks and Biological Examples

Concerning the Orthology module, the already available benchmarks for OrthoMCL still apply, except for the all-vs-all protein search step where the usage of USEARCH provides considerable time gains. For example, the complete pipeline can be

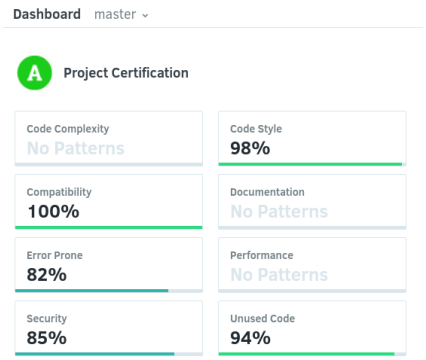

Figure 4.3. Project quality certification of TriFusion by Codacy.

executed with 10 fungal proteomes in under one hour on a regular laptop, whereas the same data set with the original OrthoMCL pipeline took approximately 32 hours.

For the Process module, a comprehensive benchmark table of all operations across a variety of data sets is hosted and continuously updated on GitHub (https://github.com/OdiogoSilva/TriFusion/wiki/Benchmarks; Table A.3.1). We tested 11 data sets that range between 1-52k files, 29-376 taxa and 17Mb-567Mb. Regardless of the data set or operation, TriFusion seldom required more than 1 minute to process it and rarely required more than 100Mb of RAM. For instance, the concatenation of 3 093 alignments with 376 taxa (~ 570Mb) can be complete in 45 seconds using less than 40Mb or RAM. We also compared TriFusion's performance to other tools that can be used for conversion and/or concatenation of alignment files across up to 10 data sets of phylogenomic size (Figure 4.4; Table A.3.1; Table A.3.1; see section 4.8.1 for details). The most important observation from these benchmarks is that TriFusion was the only tool capable of processing all tested data sets, with all other tools failing some data sets due to crashes, excessive RAM usage or timeouts. Moreover, TriFusion can be up to 3 orders of magnitude faster than the alternatives while using significantly less RAM. Therefore, the extensive code and design optimizations implemented in TriFusion allow it to provide unparalleled capabilities of processing the ever growing phylogenomic data sets. Tasks that were previously very time-consuming and had to be delegated to server-grade machines can now be executed in a matter of seconds in off-the-shelf computers.

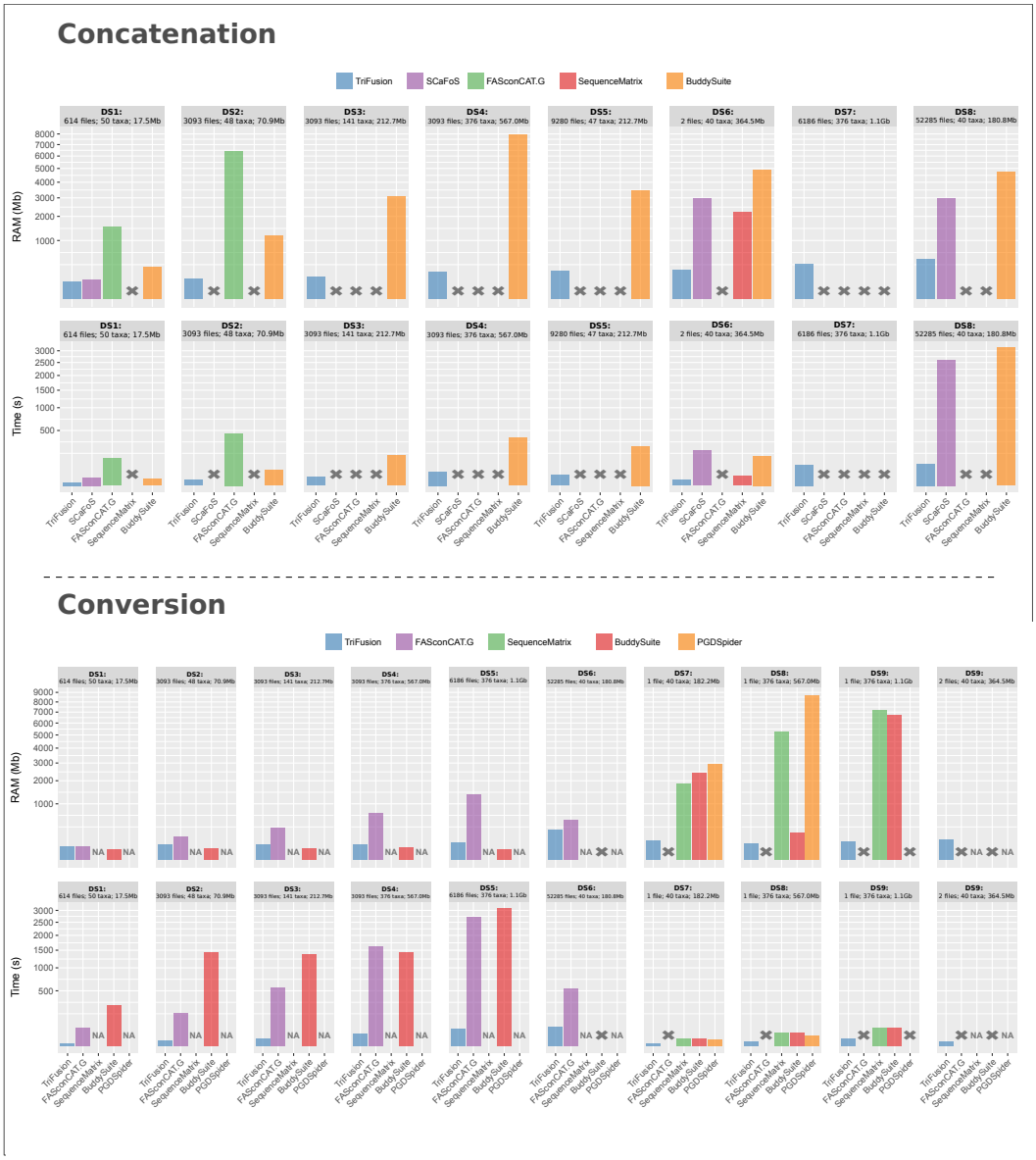

Figure 4.4. Conversion and concatenation benchmarks of TRIFUSION in comparison to other 6 software tools. The "X" symbol represents executions that the corresponding software could not conclude for that data set. "NA" represents particular cases where the conversion of many files is not supported by the corresponding software. The scale of the y-axis is square-root transformed.

4.7 Availability

TRIFUSION is freely available under the GNU General Public Licence version 3.0 and the source code is hosted on GitHub (https://github.com/ODiogoSilva/TriFusion). The software is available and easy to install on the major operating systems. In TRIFUSION's website (http://odiogosilva.github.io/TriFusion/), installers and binary packages are provided for Windows, MacOS and Linux. In addi-

tion, TriFusion can be installed in Debian-based Linux systems via Personal Package Archives (PPAs) and on ArchLinux systems using PKGBUILDS. The `trifusion` package is also available in the Python Package Index (PyPi) and instructions on how to install it directly from source are provided. There is also a comprehensive user guide available as well as a dedicated readthedocs web page with several tutorials covering all modules (http://trifusion.readthedocs.io/). A reference to TriFusion's API is also documented and available.

4.8 Supplementary data

4.8.1 Benchmarks with third-party software

TriFusion offers a wide range of features and functionalities to deal with proteome and sequence alignment data that are not available in currently existing tools. However, some basic operations like conversion and concatenation of alignment sequence data are also performed by several other programs. In this section, the performance of TriFusion was compared with these existing tools for these operations and using several phylogenomic data sets with different characteristics. The results and a brief discussion of the results are provided in section 4.6.

4.8.1.1. Hardware

Benchmarks were performed on a Intel i7-7500 @ 2.70Ghz; Intel HD620; NVMe SSD using ArchLinux.

4.8.1.2. Measurements

Measurements of run time were performed using GNU `time` program:

```
time TriSeq -in <input_files> -c
```

Measurements of peak RAM usage were performed using the `memusg` script (https://gist.github.com/netj/526585):

```
memusg TriSeq -in <input_files> -c
```

4.8.1.3. Data sets

Table 4.8.1. Overview of tests data sets for the *concatenation* benchmarks.

Data set	# of Files	# of Taxa	Bases/Residues
DS1	614	50	17.536Mb
DS2	3093	48	72.385Mb
DS3	3093	141	212.630Mb
DS4	3093	376	567,013Mb
DS5	9280	47	212.630Mb
DS6	2	40	364.531Mb
DS7	6186	376	1134.026Mb
DS8	52285	40	180.802Mb

Table 4.8.2. Overview of the tests data sets for the *conversion* benchmarks.

Data set	# of Files	# of Taxa	Bases/Residues
DS1	614	50	17.536Mb
DS2	3093	48	72.385Mb
DS3	3093	141	212.630Mb
DS4	3093	376	567,013Mb
DS5	6186	376	238.638Mb
DS6	52285	40	180.802Mb
DS7	1	40	182.225Mb
DS8	1	376	567.013Mb
DS9	1	376	1134.026Mb
DS10	2	40	364.531Mb

4.8.1.4. TriFusion

Concatenation Basic concatenation of multiple alignment files into Nexus format was performed with the following command:

```
TriSeq -in <input_files> -o <output_file>
```

Conversion Conversion of each input alignment into Nexus format was performed with the following command:

```
TriSeq -in <input_files> -c
```

4.8.1.5. SCaFoS

Concatenation Basic concatenation of multiple files into Nexus format was performed with two commands. First the generation of the OTU file:

```
scafos in=<path_to_input_dir> out=<output_dir>
```

Then, the concatenation was performed by providing the OTU file:

```
scafos in=<path_to_input_dir> out=<output_dir> otu=<otu_file>
```

Notes on SCaFoS execution When using input alignments in Fasta format, the extension of the files must be `.fasta`. Using files with another extensions, such as `.fas` yields an error:

```
None correct input file in the directory
```

It was also not possible to execute the script with several Nexus input files, which yielded the following error:

```
Number of found sequences (1) is different than predeclared
  ↪ value (40) in file
```

Even though the number of sequences and the declared value in the Nexus preamble were correct.

4.8.1.6. FASconCAT-G

Concatenation Basic concatenation of multiple input alignments into Nexus format was performed with the following command inside the directory containing the input alignment files:

```
FASconCAT-G_v1.02.pl -a -n -s
```

Conversion Basic conversion of multiples input alignments into nexus format was performed with the following command inside the directory containing the input alignment files:

```
FASconCAT-G_v1.02.pl -o -a -n -s
```

Notes on FASconCAT-G execution When using input alignments in Fasta format, the extension of the files must be `.fas`. Otherwise, the program would return the error:

```
!FILE-ERROR!: Cannot READ IN infile(s)!
```

4.8.1.7. SequenceMatrix

Concatenation The SequenceMatrix GUI application was executed with the following command:

```
java -jar -Xmx14500M -Xms256M SequenceMatrix.jar
```

Concatenation was performed by dragging and dropping input alignment files into the application window and then exporting sequences as a Nexus file. However, with the exception of one data set, the test data sets could not be imported into SequenceMatrix within a period of 60 minutes. At this point, having shown no noticeable progress, the application had to be forcibly terminated. This did not seem to be RAM related, since RAM usage was generally low and did not increase during this period.

Conversion of single files While SequenceMatrix was not developed as a conversion tool for multiple alignment files, it can be used to convert single files. The conversion of single files was performed in the same way as the concatenation.

4.8.1.8. PGDSpider

Conversion of single files PGDSpider was not developed as a conversion tool for multiple alignment files, but rather of single files. The authors also state in the program's website that it was not developed to convert very large High Throughput Sequencing (HTS) files. Nevertheless, we included it in our benchmarks of single file conversion using the CLI version.

First a `spid` file was created using the GUI version with the instruction for Fasta to Nexus conversion. Then, the conversion of single Fasta files into a Nexus output was performed with the following command:

```
java -Xmx14000M -Xms250M -jar PGDSpider2-cli.jar -inputfile
↪ <input_file> -inputformat FASTA -outputfile <output_file>
↪ -outputformat NEXUS -spid <spid.file>
```

4.8.1.9. BuddySuite

Concatenation Concatenation of individual alignment files into Nexus format was performed with the following command:

```
alignbuddy *.fas -cta ".*|..." -o nexus > concatenation.nex
```

Conversion Conversion of single alignment files into Nexus format was performed with the following command:

```
alignbuddy file.fas -uc -o nexus > file.nex
```

The conversion of multiple alignment files into Nexus format was performed inside a bash loop:

```
for i in *.fas
do alignbuddy ${i} -uc -o nexus > ${i\%fas}nex
done
```

4.9 References

Bond, S. R., K. E. Keat, S. N. Barreira, and A. D. Baxevanis (2017). „BuddySuite: Command-Line Toolkits for Manipulating Sequences, Alignments, and Phylogenetic Trees“. *Molecular Biology and Evolution* 34.6, pp. 1543–1546.

Camacho, C, G Coulouris, V Avagyan, N Ma, J Papadopoulos, K Bealer, and T. L. Madden (2009). „BLAST+: architecture and applications“. *BMC Bioinformatics* 10, p. 421.

Edgar, R. C. (2010). „Search and clustering orders of magnitude faster than BLAST“. *Bioinformatics* 26.19, pp. 2460–2461.

Emms, D. M. and S. Kelly (2015). „OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy“. *Genome Biology* 16.1, p. 157.

Enright, a. J., S Van Dongen, and C. a. Ouzounis (2002). „An efficient algorithm for large-scale detection of protein families.“ *Nucleic Acids Research* 30.7, pp. 1575–1584.

Horiike, T., R. Minai, D. Miyata, Y. Nakamura, and Y. Tateno (2016). „Ortholog-finder: A tool for constructing an ortholog data set“. *Genome Biology and Evolution* 8.2, pp. 446–457.

Iglewicz, B and D Hoaglin (1993). *How to Detect and Handle Outliers*. University of California: ASQC Quality Press, p. 87.

Kück, P. and G. C. Longo (2014). „FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies“. *Frontiers in Zoology* 11.1, p. 81.

Li, N. and M. Stephens (2003). „Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data“. *Genetics* 165.4, pp. 2213–2233.

Lischer, H. E. L. and L. Excoffier (2012). „PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs“. *Bioinformatics* 28.2, pp. 298–299.

Roure, B., N. Rodriguez-Ezpeleta, and H. Philippe (2007). „SCaFoS: a tool for Selection, Concatenation and Fusion of Sequences for phylogenomics“. *BMC Evolutionary Biology* 7, S2.

Vaidya, G., D. J. Lohman, and R. Meier (2011). „SequenceMatrix: Concatenation software for the fast assembly of multi-gene datasets with character set and codon information“. *Cladistics* 27.2, pp. 171–180.

Wu, M., S. Chatterji, and J. a. Eisen (2012). „Accounting For Alignment Uncertainty in Phylogenomics“. *PLoS ONE* 7.1, e30288.

5 Final Remarks

This is a time of a remarkable transition in modern biology. During the development of this PhD project, High Throughput Sequencing (HTS) technologies have gone from a relatively novel approach to a routine affair in biological research. Indeed, the democratization of sequencing has allowed for scientific research to go beyond what was previously possible for non-model organisms and to tackle new challenges that were inconceivable thus far. These technologies are now an integrative part of many biological fields and Illumina alone, the current dominant platform on the sequencing market, has developed more than 150 different experimental strategies that make use of HTS instruments (Muir et al., 2016). However, as a consequence, the field is now suffering from an issue of overabundance of data, with more than 14 000 genomes deposited in the US National Center for Biotechnology Information (NCBI) genome repositories and with no signs of deceleration (Goodwin et al., 2016).

With sequencing prices falling rapidly and with the continuous expansion of sequence databases, the bottleneck currently lies on the translation of this breadth of data into biological insights. This has proven to be a challenge from both analytical and infrastructure perspectives and continuous advances in bioinformatics and computational biology are required just to keep up with the pace (Pavey et al., 2012; Muir et al., 2016). Moreover, the lack of gold standards or even of an agreement as to which are the best strategies for processing and interpreting massive data sets are a sign that the "omics" fields are still a few steps from maturity, even though advances are being made at an appalling pace. Nowadays, researchers that want to make use of genome-scale approaches to address biological questions are faced with a myriad of non-trivial challenges at the analytical level. What are the best approaches for cleaning data before starting the assembly of the data set? Which tools should be used to assemble the data set? How does one measure the quality of the assembled data and decide how to best proceed with the analysis? Do we use

established statistical approaches developed before the widespread use of genome-scale data, or do we attempt to use new experimental approaches? Are the available tools sufficient to achieve our goals or is it necessary to develop new software? It was amidst the constant advances in sequencing technologies combined with the unwavering emergence of new technical and analytical challenges that this thesis was developed, with a focus on the application of genomic approaches to study fungal pathogens.

Fungi (including Oomycetes) cause some of the most devastating plant diseases with dramatic impacts on human well-being by risking food security and causing economical losses (Giraud, 2006; Gandon et al., 2016). Despite this impact, human activities have further exacerbated the problem and fungal infectious diseases will continue to emerge at an alarming rate in the foreseeable future (Gandon et al., 2016). To mitigate, prevent and reverse this trend, one aspect seems to have become clear: Eco-evolutionary principles will need to be included in the delineation of measures to control infectious diseases and this will require a deeper understanding of the genetic mechanisms of fungal pathogenicity and the evolutionary history of pathogen populations (Mcdonald, 2015; Zhan et al., 2015; Plissonneau et al., 2017).

In this work, the informative potential of genome-scale data and the new research venues that were unlocked by the HTS revolution were leveraged for further our knowledge on a destructive group of fungal pathogens, the rusts, with a particular focus on coffee rust.

5.1 Conclusions and main contributions

5.1.1 Phylogenomics

In chapter 2 of this thesis, we investigated the role of positive selection at the phylogenetic root of the rust fungi (Pucciniales) acting on genes that were shared and conserved across a broad range of the Basidiomycota. When this work was initiated, the genomic resources of the rust fungi were just starting to be released and

there was a large and coordinated effort to mine those genomes for exclusive features that could explain the unique biotrophic life-style and the genomic determinants of rust pathogenicity. However, the role of adaptive variation on shared genetic material remained unexplored. To tackle this issue while taking full advantage of the recent genomic resources, we developed an entire phylogenomic pipeline from scratch that made use of several software tools considered state of the art (this pipeline also laid the foundation and motivation for TriFusion's development in chapter 4). This resulted in the assembly of one of the largest data matrices for the Basidiomycota and fungi in general, and this data set was then used to investigate genome-wide patterns of positive selection at the evolutionary origin of the rust fungi. Such genome-scans were still uncommon due to the scarcity of public complete genomes and only two had been performed in the fungal kingdom (Aguileta et al., 2010; Aguileta et al., 2012). This meant that such a novel approach could bring fresh insights into the problem that was being tackled but, at the same time, there would be a lack of bioinformatic support for the task.

To overcome this constraint, an automated genome scanning pipeline for the detection of positive selection on a specific phylogenetic branch was developed and the results revealed that the role of positive selection on shared genetic variation was much higher than previous genome-scans had shown. Signals of positive selection were found on 19.6 - 33.3% of the orthologous genes screened. This considerable signal of natural selection was attributed to both a biologically relevant signal and the fact that a fairly long and basal phylogenetic branch was being scanned, which tends to increase the power to detect selection. Moreover, our new approach of profiling the specific amino acids under positive selection according to specified classes revealed a new layer of information that was never explored in a context of a genome-wide scan. With this information we discovered several functional classes enriched for positively selected genes in our data set, identifying processes that could be important to the origin of the rust fungi and, consequently, of their unique life-style. Most of these functional classes were related to nutrient metabolism and uptake, which can be explained by the changes that these fungi underwent while transitioning to an obligate biotrophic life-style. Moreover, several positively selected genes involved in secondary metabolism were also detected, which are described

in the literature as triggers of host defense, and their genetic diversification may be a mechanism to avoid host recognition.

However, the most intriguing result of this study was the discovery that a substantial proportion (15-24%) of the amino acids with a signal of positive selection were actually conserved across all or most species of our data set. This occurred for amino acids that can be encoded by a wide range of codons, such as Serine which is translated by both TCN and AGY codons. In these cases, the transitions between these codons may require 2-3 non-synonymous mutations just to revert to the original amino acid. Furthermore, we also found that the codon usage for these amino acids was different between rust and non-rusts. In this thesis, several hypothesis were proposed to explain these results without resorting to the role of positive selection. These include stochastic variation maintained by purifying/stabilizing selection, simultaneous double nucleotide substitutions and the occurrence of a mutational bias. In the end, I am still more convinced that this represents a biologically meaningful indication that codon usage has a much more preponderant role in the evolution of organisms, or at least of rust fungi, and warrants future investigation.

There are several reasons to explain how this phenomenon would go unnoticed, but two stand out in particular. First, genome-scans for positive selection are still not very common due to the amount of work involved and the profiling of positively select amino acids, which allowed us to uncover it, was not done before. Second, the discovery of conserved selected amino acids requires the genome scan to be performed on a taxonomically broad data set to evaluate the degree of amino acid conservation. However, the majority of the studies performing genome scans or testing for the presence of positive selection are undertaken with a limited sampling that is usually the minimum required to address the problem. In fact, without our effort to include the largest possible number of representatives of the Basidiomycota, this phenomenon would probably remain overlooked.

5.1.2 Population genomics

In chapter 3 of this thesis, the focus was shifted to the main coffee pathogen, *Hemileia vastatrix* causing Coffee Leat Rust (CLR) in coffee crops worldwide. Due to the occurrence of a series of cluster outbreaks of CLR across Latin America, the re-emergence of *H. vastatrix* triggered emergency actions in the main coffee producing nations and reached the status of natural disaster in some regions (Cressey, 2013; Talhinhas et al., 2017). In this work, we investigated *H. vastatrix*'s populations at the genomic level using RAD sequencing, in an effort to translate population genomic insights into useful recommendations for disease control. The results revealed that this pathogen, which was previously though to be a single unstructured species able to infect multiple coffee species, was in fact most likely a complex of cryptic species with marked host tropism. This result is in line with the growing notion that cryptic diversity is very high within the Pucciniales order (Liu and Hambleton, 2013; Mctaggart et al., 2015; Aime et al., 2017). Within the three divergent lineages found, the most recent among the complex seems to be a "domesticated" and recombinant group of isolates that exclusively infects the most economically important coffee species, *C. arabica* and tetraploid inter-specific hybrids, and appear to have resulted from a host-shift from populations adapted to diploid coffee hosts. Indeed, speciation by host-shift, from taxonomically close hosts, has been shown to be a favored evolutionary venue in rust fungi (Aime, 2006; Mctaggart et al., 2015; Aime et al., 2017). In the case of *H. vastatrix*, this resulted in a paradigm change on how we look at its epidemics and will have a significant impact on future research on the pathogen. For instance, the long held notion that diploid *C. canephora* plants are more resistant to *H. vastatrix* seems to be a reflection of the specialization of the most epidemiological important group of *H. vastatrix* to tetraploid coffee hosts and not of inherent resistance factors. Diploid coffee plants seem to be equally susceptible when infected with *H. vastatrix* lineages adapted to these hosts. Nevertheless, the resistance factors of diploid coffee hosts have been the basis of coffee breeding programs worldwide, via introgression from natural hybrids between *C. arabica* and *C. canephora* (known as "Híbrido de Timor"), the major source of coffee rust resistance. Therefore, the discovery that hybridization and introgression can also occur between different lineages of *H. vastatrix* adapted to different sets

of hosts raises the alarming possibility that virulence factors may be exchanged. This also warns researchers and coffee breeders in general that the highest risk for the generation of new hyper-virulent strains may come from plantations and nurseries where diploid and tetraploid coffee hosts are maintained in close proximity. Overall, the results from this study provide a striking example of both the power of population genomics in unraveling the complex and often elusive evolutionary history of plant pathogen populations, and how the dynamics of these populations have been heavily influenced by agro-ecosystems. Moreover, these findings will need to be taken into account in future research of the pathogen, particularly when it comes to the investigation of its pathogenicity and the association of molecular markers to pathotypes.

5.1.3 Software development

Chapter 4 of this thesis describes a novel software application that was developed from the ground up during the period of this thesis, named TriFusion. The main motivation and the initial code base for this project came from the development of the phylogenomic pipeline performed in chapter 2, and from the lack of efficient tools in the field, user-friendly or otherwise. Since I had already gathered substantial knowledge in the intricacies of building phylogenomic pipelines and had a reasonable grasping of computer programming, I decided to begin the project of building an actual graphical desktop application to facilitate the life of other researches in the field. Little did I know, however, the amount of time and effort that would be required to build such a cross-platform application, including a graphical interface, Application Programming Interface (API) and user documentation, a comprehensive set of tests, benchmarks, installers and packages for the major operating systems, a web page and animated tutorials. Nevertheless, from an early set of scripts, TriFusion turned into a fully fledged and feature rich desktop and command line application for gathering, processing and visualization of massive data sets that is accessible to a wide range of researchers regardless of bioinformatics knowledge. This application proved to be an essential companion tool throughout the development of the tasks in this thesis, and in further work collaborations, when processing large sets of sequence data in phylogenomic and population genomic studies. It allows the

search of ortholog genes across multiple proteomes with just a few clicks, instead of the complicated and laborious processes of running the OrthoMCL scripts and subsequent filtering. It also greatly outperforms all other available tools in the most common tasks of sequence conversion and concatenation, while providing a whole range of new functionalities to deal with this kind of data. Therefore, we believe that this software will represent a major contribution to the field and that it might have a broad impact on the scientific community. As a final ironic note, even though the graphical user interface was the most labored element of the application and one of the greatest selling points, its main developer remains an affix fan of command line interfaces.

5.2 Future perspectives

Overall, the results of this thesis demonstrate the power of applying phylogenomic and population genomic approaches to investigate and better understand the origin and emergence of fungal pathogens. However, the most exciting outcomes of these approaches are often the novel questions and challenges that come to light. In the context of phylogenomics, the ever growing number of available genomes and the increasing sophistication of analytical tools will allow for the exploration of genome-wide patterns of natural selection to be unraveled with greater detail. In this thesis we only explored how positive selection shaped the evolution of genes at a particular branch in the Basidiomycota tree, and naturally this represents only a snapshot of natural selection's big picture. However, it is now possible to conduct these genome-scans for a larger set of genomes that cover a broader taxonomic range and evaluate how selection acts on contiguous branches of the tree of life. This will allow researchers to understand how natural selection shapes organism evolution both along the genomes and through time. Perhaps more exciting will be the future research on the apparent paradox of conserved amino acids under positive selection resulting in different codon preferences, and how the use of different codons provide an advantage in different taxa. Whether this signal will prove to be the result of positive selection or other evolutionary forces that does not implicate fitness or adaptive changes, remains an open question. Nevertheless, this

phenomenon has left a significant footprint on the evolution of the rust fungi. The extent of this phenomenon in other taxonomic groups is currently unknown, but it would be interesting to investigate whether it is a general occurrence or a rare event that is specific to some groups of organisms.

In the context of population genomics, the application of genomic-scale data to investigate the evolutionary history of the coffee rust pathogen resulted in a paradigm shift in our understanding of the pathogen and its epidemiology. Now that we have a much more detailed and informative perspective of *H. vastatrix*, research on this pathogen can proceed in several interesting ways. Given the existence of hybridization between fungal lineages that have very distinct pathogenic abilities, it will be important to understand whether the introgressed loci comprise virulence factors that would allow the pathogen to overcome host resistance. Research on rust fungi is still complicated due to the biotrophic lifestyle but genomic approaches coupled with a suitable study design would make this investigation possible. Another important investigation route would be the association of genetic markers to particular pathotypes, taking the cryptic species complex structure into consideration. An essential requirement of genome-wide association studies is a solid understanding of the population structure underlying the sampling, and this work provides a basis for such endeavor. However, I also stress that the search for specific SNPs associated with any particular *H. vastatrix* pathotype will not be a simple task, particularly for the most epidemiological important group. Using a reduced representation library approach to generate SNP data for *H. vastatrix* revealed a reduced number of shared polymorphism among isolates of the *C3* group, and this provides little ground to perform association studies. Given the low genetic variability of this group, the quest for this kind of association studies will probably require a higher sequencing effort.

As for the software development and general bioinformatics that were performed during this project, it seems clear to me that these skills are essential for the successful execution of genomic projects. From the perspective of a software user, this is a fast paced field and it is often difficult to catch up with everything that is released and of interest to our research. It is not uncommon to spend a considerable amount of time learning the ropes of a promising piece of software, only to see it being superseded

shortly after by a new version or program. Personally, this taught me the huge value that good documentation adds to any software. From the perspective of a software developer, it is also an incredibly challenging task to build and maintain usable and relevant software in the field. Even though I made many other code contributions to the field during this thesis, TriFusion was and still is my most cherished project; the one where I tried to invest and develop the good software development practices that I learned along the way. One of those practices was to build and document the application in a way that it would be easy to contribute and expand with new features in the future. Only time will tell whether TriFusion will be future-proof, but in any case and considering the work of this thesis as a whole, nothing will take away the incredible learning journey that this project was.

5.3 References

Aguileta, G, J Lengelle, S Marthey, H Chiapello, F Rodolphe, A Gendrault, R Yockteng, E Vercken, B Devier, M. C. Fontaine, et al. (2010). „Finding candidate genes under positive selection in Non-model species: examples of genes involved in host specialization in pathogens." *Molecular Ecology* 19.2, pp. 292–306.

Aguileta, G., J. Lengelle, H. Chiapello, T. Giraud, M. Viaud, E. Fournier, F. Rodolphe, S. Marthey, A. Ducasse, A. Gendrault, et al. (2012). „Genes under positive selection in a model plant pathogenic fungus, Botrytis." *Infection, Genetics and Evolution* 12.5, pp. 987–996.

Aime, M. C. (2006). „Toward resolving family-level relationships in rust fungi (Uredinales)". *Mycoscience* 47.3, pp. 112–122.

Aime, M. C., A. R. McTaggart, S. J. Mondo, and S. Duplessis (2017). „Phylogenetics and Phylogenomics of Rust Fungi". In: *Advances in Genetics*. Vol. 100, pp. 267–307.

Cressey, D. (2013). „Coffee rust regains foothold Researchers marshal technology in bid to thwart fungal outbreak in Central America". *Nature* 493.7434, p. 587.

Gandon, S., T. Day, C. J. E. Metcalf, and B. T. Grenfell (2016). „Forecasting Epidemiological and Evolutionary Dynamics of Infectious Diseases". *Trends in Ecology & Evolution* 31, pp. 776–788.

Giraud, T (2006). „Selection against migrant pathogens: the immigrant inviability barrier in pathogens." *Heredity* 97, pp. 316–318.

Goodwin, S., J. D. McPherson, and W. R. McCombie (2016). „Coming of age: ten years of next-generation sequencing technologies". *Nature Reviews Genetics* 17.6, pp. 333–351.

Liu, M. and S. Hambleton (2013). „Laying the foundation for a taxonomic review of Puccinia coronata s.l. in a phylogenetic context". *Mycological Progress* 12.1, pp. 63–89.

Mcdonald, B. A. (2015). „How can research on pathogen population biology suggest disease management strategies? The example of barley scald (Rhynchosporium commune)“. *Plant Pathology* 64.5, pp. 1005–1013.

Mctaggart, A. R., C Dougsa-ard, A. Geering, M. C. Aime, and R. G. Shivas (2015). „A co-evolutionary relationship exists between Endoraecium (Pucciniales) and its Acacia hosts in Australia“. *Persoonia* 35, pp. 50–62.

Muir, P., S. Li, S. Lou, D. Wang, D. J. Spakowicz, L. Salichos, J. Zhang, G. M. Weinstock, F. Isaacs, J. Rozowsky, et al. (2016). „The real cost of sequencing: scaling computation to keep pace with data generation“. *Genome Biology* 17.1, p. 53.

Pavey, S. A., L. Bernatchez, N. Aubin-Horth, and C. R. Landry (2012). „What is needed for next-generation ecological and evolutionary genomics?“ *Trends in Ecology & Evolution* 27.12, pp. 673–678.

Plissonneau, C., J. Benevenuto, N. Mohd-Assaad, S. Fouché, F. E. Hartmann, and D. Croll (2017). „Using Population and Comparative Genomics to Understand the Genetic Basis of Effector-Driven Fungal Pathogen Evolution“. *Frontiers in Plant Science* 8.February, pp. 1–15.

Talhinhas, P., D. Batista, I Diniz, A Vieira, D. Silva, A Loureiro, S Tavares, A. Pereira, H. Azinheira, L Guerra-Guimarães, et al. (2017). „Pathogen profile The coffee leaf rust pathogen Hemileia vastatrix : one and a half centuries around the tropics“. *Molecular Plant Pathology* 18.8, pp. 1039–1051.

Zhan, J., P. H. Thrall, J. Papaïx, L. Xie, and J. J. Burdon (2015). „Playing on a Pathogen's Weakness: Using Evolution to Guide Sustainable Plant Disease Control Strategies“. *Annual Review of Phytopathology* 53.1, pp. 19–43.

A

Appendix

A.1 Genomic patterns of positive selection at the origin of rust fungi

A.1.1 Tables

Table A.1.1. Details on the genomic and EST data used on the present study.

Species	Data type	Taxonomy				Data set		
		Phylum	Subphylum	Class	Order	Protein number	Total length	Average length
Agaricus bisporus	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Agaricales	10438	4514068	432,51
Auricularia delicata	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Auriculariales	23577	9238049	391,84
Bjerkandera adusta	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Polyporales	15473	6379989	412,36
Ceriporiopsis subvermispora	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Polyporales	12125	5249589	432,99
Coniophora puteana	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Boletales	13761	6116895	444,54
Coprinopsis cinerea	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Agaricales	13342	6283763	471,01
Cryptococcus gattii	Genome	Basidiomycota	Agaricomycotina	Tremellomycetes	Tremellales	6210	3254637	524,18
Cryptococcus neoformans	Genome	Basidiomycota	Agaricomycotina	Tremellomycetes	Tremellales	6967	3637167	522,13
Dacryopinax sp.	Genome	Basidiomycota	Agaricomycotina	Dacrymycetes	Dacrymycetales	10242	4210446	411,14
Dichomitus squalens	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Polyporales	12290	5225671	425,23
Fomitiporia mediterranea	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Hymenochaetales	11333	4919745	434,15
Fomitopsis pinicola	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Polyporales	14724	5919823	402,08
Ganoderma sp.	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Polyporales	12910	5599385	433,76
Gloeophyllum trabeum	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Gloeophyllaceae	11846	5059364	427,13
Heterobasidion annosum	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Russulales	13405	5154689	384,56
Laccaria bicolor	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Agaricales	23132	8353592	361,14
Melampsora laricis-populina	Genome	Basidiomycota	Pucciniomycotina	Pucciniomycetes	Pucciniales	16694	6410403	384,02
Paxillus involutus	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Boletales	17968	6494473	361,47
Phanerochaete carnosa	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Polyporales	13937	5428644	389,54
Phanerochaete chrysosporium	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Polyporales	10048	4643668	462,19
Phlebia brevispora	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Polyporales	16170	6567691	406,19

Continued on next page

Table A.1.1. Details on the genomic and EST data used on the present study.

Species	Data type	Taxonomy				Data set		
		Phylum	Subphylum	Class	Order	Protein number	Total length	Average length
Phlebiopsis gigantea	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Polyporales	11891	4964705	417,55
Pleurotus ostreatus	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Agaricales	12330	5226749	423,94
Postia placenta	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Polyporales	9113	4000497	439,04
Puccinia graminis f. sp. Tritici	Genome	Basidiomycota	Pucciniomycotina	Pucciniomycetes	Pucciniales	15979	6437350	402,89
Puccinia triticina	Genome	Basidiomycota	Pucciniomycotina	Pucciniomycetes	Pucciniales	11630	4745807	408,10
Punctularia strigosozonata	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Corticiales	11538	5154085	446,74
Schizophyllum commune	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Agaricales	14652	6550036	447,07
Serpula lacrymans	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Boletales	14495	4744658	327,35
Serpula lacrymans S7.9	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Boletales	16257	5179401	318,61
Sporobolomyces roseus	Genome	Basidiomycota	Pucciniomycotina	Microbotryomycetes	Sporidiobolales	5536	3173802	573,41
Stereum hirsutum	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Russulales	14072	6537368	464,60
Trametes versicolor	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Polyporales	14296	6122306	428,28
Tremella mesenterica	Genome	Basidiomycota	Agaricomycotina	Tremellomycetes	Tremellales	8313	3929320	472,73
Ustilago maydis	Genome	Basidiomycota	Ustilaginomycotina	Ustilaginomycetes	Ustilaginales	6522	4069824	624,11
Wallemia sebi	Genome	Basidiomycota	incertae sedis	Wallemiomycetes	Wallemiales	5284	2268239	429,35
Wolfiporia cocos	Genome	Basidiomycota	Agaricomycotina	Agaricomycetes	Polyporales	12746	5290821	415,13
Armillaria tabescens	EST	Basidiomycota	Agaricomycotina	Agaricomycetes	Agaricales	57	2126	37,30
Cryptococcus laurentii	EST	Basidiomycota	Agaricomycotina	Tremellomycetes	Tremellales	1200	49383	41,15
Fomitopsis palustris	EST	Basidiomycota	Agaricomycotina	Agaricomycetes	Polyporales	535	33072	61,82
Ganoderma lucidum	EST	Basidiomycota	Agaricomycotina	Agaricomycetes	Polyporales	2517	160874	63,91
Hemileia vastatrix	EST	Basidiomycota	Pucciniomycotina	Pucciniomycetes	Pucciniales	1311	332994	254,58
Lentinula edodes	EST	Basidiomycota	Agaricomycotina	Agaricomycetes	Agaricales	1970	97881	49,69

Continued on next page

Table A.1.1. Details on the genomic and EST data used on the present study.

Species	Data type	Taxonomy				Data set		
		Phylum	Subphylum	Class	Order	Protein number	Total length	Average length
Leucosporidium scottii	EST	Basidiomycota	Pucciniomycotina	Microbotryomycetes	Leucosporidiales	1311	73103	55,76
Melampsora medusae f. sp. deltoidis	EST	Basidiomycota	Pucciniomycotina	Pucciniomycetes	Pucciniales	275	13265	48,24
Melampsora medusae f. sp. tremuloidis	EST	Basidiomycota	Pucciniomycotina	Pucciniomycetes	Pucciniales	317	11769	37,13
Melampsora occidentalis	EST	Basidiomycota	Pucciniomycotina	Pucciniomycetes	Pucciniales	452	21870	48,38
Microbotryum violaceum	EST	Basidiomycota	Pucciniomycotina	Microbotryomycetes	Microbotryales	1916	91847	47,94
Moniliophthora perniciosa	EST	Basidiomycota	Agaricomycotina	Agaricomycetes	Agaricales	471	21725	46,13
Phakopsora pachyrhizi	EST	Basidiomycota	Pucciniomycotina	Pucciniomycetes	Pucciniales	1239	174528	140,86
Phellinidium sulphurascens	EST	Basidiomycota	Agaricomycotina	Agaricomycetes	Hymenochaetales	413	23148	56,05
Pisolithus microcarpus	EST	Basidiomycota	Agaricomycotina	Agaricomycetes	Boletales	214	10046	46,94
Pisolithus tinctorius	EST	Basidiomycota	Agaricomycotina	Agaricomycetes	Boletales	161	7958	49,43
Puccinia striiformis f. sp. tritici	EST	Basidiomycota	Pucciniomycotina	Pucciniomycetes	Pucciniales	324	15405	47,55
Suillus luteus	EST	Basidiomycota	Agaricomycotina	Agaricomycetes	Boletales	149	3954	26,54
Taiwanofungus camphoratus	EST	Basidiomycota	Agaricomycotina	Agaricomycetes	Polyporales	1608	76415	47,52
Uromyces appendiculatus	EST	Basidiomycota	Pucciniomycotina	Pucciniomycetes	Pucciniales	1033	46671	45,18
Uromyces viciae-fabae	EST	Basidiomycota	Pucciniomycotina	Pucciniomycetes	Pucciniales	173	6420	37,110
OUTGROUP								
Saitoella complicata	Genome	Ascomycota	Taphrinomycotina	incertae sedis	incertae sedis	7034	3023387	429,89
Yarrowia lipolytica	Genome	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales	6448	3194468	495,5
Lipomyces starkeyi	Genome	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales	9493	3498639	368,59
Mycosphaerella fijiensis	Genome	Ascomycota	Pezizomycotina	Dothideomycetes	Capnodiales	13107	5688215	434,02
Cochliobolus heterostrophus	Genome	Ascomycota	Pezizomycotina	Dothideomycetes	Pleosporales	13336	5795028	434,57
Botrytis cinerea	Genome	Ascomycota	Pezizomycotina	Leotiomycetes	Helotiales	16448	5611312	341,18

Continued on next page

Table A.1.1. Details on the genomic and EST data used on the present study.

Species	Data type	Taxonomy				Data set		
		Phylum	Subphylum	Class	Order	Protein number	Total length	Average length
Colletotrichum higginsianum	Genome	Ascomycota	Pezizomycotina	Sordariomycetes	Glomerellales	16150	6001965	371,66
Nectria haematococca	Genome	Ascomycota	Pezizomycotina	Sordariomycetes	Hypocreales	15707	7661162	487,79
Aspergillus carbonarius	Genome	Ascomycota	Pezizomycotina	Eurotiomycetes	Eurotiales	11624	5431323	467,29

Table A.1.2. Summary statistics about missing data information and average gene length for the `basidioPAML` data set.

Species	MGenes [a]	MGenes (%) [a]	Tgenes [b]	Mdata [c]	Gaps	TMdata[d]	TMdata (%) [d]	Tchars[e]	AGL[f]
Agaricus bisporus	65	6,46%	1 006	36 543	52 614	89 157	11,18%	797 307	752,55
Auricularia delicate	48	4,77%	1 006	29 700	54 588	84 288	10,57%	797 307	744,28
Bjerkandera adusta	14	1,39%	1 006	8 599	56 343	64 942	8,15%	797 307	738,28
Ceriporiopsis subvermispora	93	9,24%	1 006	56 832	48 813	105 645	13,25%	797 307	757,58
Coniophora puteana	63	6,26%	1 006	38 175	50 424	88 599	11,11%	797 307	751,55
Coprinopsis cinerea	23	2,29%	1 006	11 946	62 796	74 742	9,37%	797 307	735,07
Cryptococcus gattii	57	5,67%	1 006	46 686	73 074	119 760	15,02%	797 307	713,99
Cryptococcus neoformans	46	4,57%	1 006	43 044	54 828	97 872	12,28%	797 307	728,59
Dacryopinax sp	53	5,27%	1 006	38 274	56 718	94 992	11,91%	797 307	736,96
Dichomitus squalens	44	4,37%	1 006	28 041	51 462	79 503	9,97%	797 307	746,16
Fomitiporia mediterranea	36	3,58%	1 006	24 825	51 363	76 188	9,56%	797 307	743,42
Fomitopsis pinicola	20	1,99%	1 006	14 091	57 063	71 154	8,92%	797 307	736,47
Ganoderma sp	15	1,49%	1 006	11 715	54 726	66 441	8,33%	797 307	737,50
Gloeophyllum trabeum	14	1,39%	1 006	11 565	57 042	68 607	8,60%	797 307	734,58
Laccaria bicolor	16	1,59%	1 006	10 785	56 199	66 984	8,40%	797 307	737,70
Melampsora laricis-populina	14	1,39%	1 006	13 701	73 902	87 603	10,99%	797 307	715,45
Paxillus involutus	25	2,49%	1 006	15 261	66 030	81 291	10,20%	797 307	729,90
Phanerochaete carnosa	91	9,05%	1 006	52 236	60 678	112 914	14,16%	797 307	747,98
Phanerochaete chrysosporium	64	6,36%	1 006	41 733	64 977	106 710	13,38%	797 307	733,30
Phlebia brevispora	21	2,09%	1 006	12 027	61 305	73 332	9,20%	797 307	735,01
Phlebiopsis gigantea	19	1,89%	1 006	10 802	59 460	70 262	8,81%	797 307	736,63
Pleurotus ostreatus	32	3,18%	1 006	21 210	58 611	79 821	10,01%	797 307	736,66

Continued on next page

Table A.1.2. Summary statistics about missing data information and average gene length for the `basidioPAML` data set.

Species	MGenes [a]	MGenes (%) [a]	Tgenes [b]	Mdata [c]	Gaps	TMdata[d]	TMdata (%) [d]	Tchars[e]	AGL[f]
Puccinia graminis f. sp. Tritici	10	0,99%	1 006	5 796	62 880	68 676	8,61%	797 307	731,60
Puccinia triticina	340	33,80%	1 006	278 697	21 123	299 820	37,60%	797 307	747,01
Punctularia strigosozonata	58	5,77%	1 006	40 194	53 739	93 933	11,78%	797 307	741,96
Schizophyllum commune	20	1,99%	1 006	12 066	59 172	71 238	8,93%	797 307	736,38
Serpula lacrymans S7.9	16	1,59%	1 006	10 653	57 180	67 833	8,51%	797 307	736,85
Sporobolomyces roseus	15	1,49%	1 006	12 406	93 063	105 469	13,23%	797 307	699,02
Stereum hirsutum	38	3,78%	1 006	23 814	51 432	75 246	9,44%	797 307	745,93
Trametes versicolor	38	3,78%	1 006	24 684	52 020	76 704	9,62%	797 307	744,42
Tremella mesenterica	182	18,09%	1 006	142 080	61 185	203 265	25,49%	797 307	726,36
Ustilago maydis	69	6,86%	1 006	49 734	48 873	98 607	12,37%	797 307	745,68
Wallemia sebi	98	9,74%	1 006	93 288	55 656	148 944	18,68%	797 307	714,09
Wolfiporia cocos	13	1,29%	1 006	10 575	56 145	66 720	8,37%	797 307	735,74
Average	52,06	5,17%		37 699,35	57 514,24	95 213,59	11,94%		736,31
Standard deviation	61,72	6,14%		50 534,16	10 649,96	45 372,45	5,69%		12,00

[a] Missing genes

[b] Total genes

[c] Missing data (in base pairs)

[d] Total missing data (in base pairs)

[e] Total characters (in base pairs)

[f] Average gene length (in base pairs)

Table A.1.3. Summary statistics about missing data information and average gene length for the `basidioPAML_Hv` data set.

Species	MGenes [a]	MGenes (%) [a]	Tgenes [b]	Mdata [c]	Gaps	TMdata[d]	TMdata (%) [d]	Tchars[e]	AGL[f]
Agaricus bisporus	65	6,46%	1 006	36 543	52 614	89 157	11,18%	797 307	752,55
Auricularia delicate	48	4,77%	1 006	29 700	54 588	84 288	10,57%	797 307	744,28
Bjerkandera adusta	14	1,39%	1 006	8 599	56 343	64 942	8,15%	797 307	738,28
Ceriporiopsis subvermispora	93	9,24%	1 006	56 832	48 813	105 645	13,25%	797 307	757,58
Coniophora puteana	63	6,26%	1 006	38 175	50 424	88 599	11,11%	797 307	751,55
Coprinopsis cinerea	23	2,29%	1 006	11 946	62 796	74 742	9,37%	797 307	735,07
Cryptococcus gattii	57	5,67%	1 006	46 686	73 074	119 760	15,02%	797 307	713,99
Cryptococcus neoformans	46	4,57%	1 006	43 044	54 828	97 872	12,28%	797 307	728,59
Dacryopinax sp	53	5,27%	1 006	38 274	56 718	94 992	11,91%	797 307	736,96
Dichomitus squalens	44	4,37%	1 006	28 041	51 462	79 503	9,97%	797 307	746,16
Fomitiporia mediterranea	36	3,58%	1 006	24 825	51 363	76 188	9,56%	797 307	743,42
Fomitopsis pinicola	20	1,99%	1 006	14 091	57 063	71 154	8,92%	797 307	736,47
Ganoderma sp	15	1,49%	1 006	11 715	54 726	66 441	8,33%	797 307	737,50
Gloeophyllum trabeum	14	1,39%	1 006	11 565	57 042	68 607	8,60%	797 307	734,58
Laccaria bicolor	16	1,59%	1 006	10 785	56 199	66 984	8,40%	797 307	737,70
Melampsora laricis-populina	14	1,39%	1 006	13 701	73 902	87 603	10,99%	797 307	715,45
Paxillus involutus	25	2,49%	1 006	15 261	66 030	81 291	10,20%	797 307	729,90
Phanerochaete carnosa	91	9,05%	1 006	52 236	60 678	112 914	14,16%	797 307	747,98
Phanerochaete chrysosporium	64	6,36%	1 006	41 733	64 977	106 710	13,38%	797 307	733,30
Phlebia brevispora	21	2,09%	1 006	12 027	61 305	73 332	9,20%	797 307	735,01
Phlebiopsis gigantea	19	1,89%	1 006	10 802	59 460	70 262	8,81%	797 307	736,63
Pleurotus ostreatus	32	3,18%	1 006	21 210	58 611	79 821	10,01%	797 307	736,66
Puccinia graminis f. sp. Tritici	10	0,99%	1 006	5 796	62 880	68 676	8,61%	797 307	731,60

Continued on next page

Table A.1.3. Summary statistics about missing data information and average gene length for the `basidioPAML_Hv` data set.

Species	MGenes [a]	MGenes (%) [a]	Tgenes [b]	Mdata [c]	Gaps	TMdata[d]	TMdata (%) [d]	Tchars[e]	AGL[f]
Puccinia triticina	340	33,80%	1 006	278 697	21 123	299 820	37,60%	797 307	747,01
Punctularia strigosozonata	58	5,77%	1 006	40 194	53 739	93 933	11,78%	797 307	741,96
Schizophyllum commune	20	1,99%	1 006	12 066	59 172	71 238	8,93%	797 307	736,38
Serpula lacrymans S7.9	16	1,59%	1 006	10 653	57 180	67 833	8,51%	797 307	736,85
Sporobolomyces roseus	15	1,49%	1 006	12 406	93 063	105 469	13,23%	797 307	699,02
Stereum hirsutum	38	3,78%	1 006	23 814	51 432	75 246	9,44%	797 307	745,93
Trametes versicolor	38	3,78%	1 006	24 684	52 020	76 704	9,62%	797 307	744,42
Tremella mesenterica	182	18,09%	1 006	142 080	61 185	203 265	25,49%	797 307	726,36
Ustilago maydis	69	6,86%	1 006	49 734	48 873	98 607	12,37%	797 307	745,68
Wallemia sebi	98	9,74%	1 006	93 288	55 656	148 944	18,68%	797 307	714,09
Wolfiporia cocos	13	1,29%	1 006	10 575	56 145	66 720	8,37%	797 307	735,74
Average	52,06	5,17%		37 699,35	57 514,24	95 213,59	11,94%		736,31
Standard deviation	61,72	6,14%		50 534,16	10 649,96	45 372,45	5,69%		12,00

[a] Missing genes

[b] Total genes

[c] Missing data (in base pairs)

[d] Total missing data (in base pairs)

[e] Total characters (in base pairs)

[f] Average gene length (in base pairs)

Table A.1.4. Summary statistics about missing data information and average gene length for the `genomic46sp_sparse` data set.

Species	MGenes [a]	MGenes (%) [a]	Tgenes [b]	Mdata [c]	Gaps	TMdata[d]	TMdata (%) [d]	Tchars[e]	AGL[f]
Agaricus bisporus	179	10,44%	1 715	91 176	45 609	136 785	16,23%	842 795	459,64
Aspergillus carbonarius	560	32,65%	1 715	307 347	46 700	354 047	42,01%	842 795	423,16
Auricularia delicate	131	7,64%	1 715	77 235	59 009	136 244	16,17%	842 795	446,05
Bjerkandera adusta	50	2,92%	1 715	36 893	41 430	78 323	9,29%	842 795	459,14
Botrytis cinerea	540	31,49%	1 715	314 300	65 199	379 499	45,03%	842 795	394,29
Ceriporiopsis subvermispora	242	14,11%	1 715	100 070	34 468	134 538	15,96%	842 795	480,83
Cochliobolus heterostrophus	504	29,39%	1 715	269 108	51 113	320 221	38,00%	842 795	431,52
Colletotrichum higginsianum	509	29,68%	1 715	308 297	55 443	363 740	43,16%	842 795	397,23
Coniophora puteana	165	9,62%	1 715	92 348	39 024	131 372	15,59%	842 795	458,98
Coprinopsis cinerea	67	3,91%	1 715	40 034	49 906	89 940	10,67%	842 795	456,83
Cryptococcus gattii	249	14,52%	1 715	142 260	64 667	206 927	24,55%	842 795	433,74
Cryptococcus neoformans	213	12,42%	1 715	122 827	55 249	178 076	21,13%	842 795	442,56
Dacryopinax sp	166	9,68%	1 715	118 716	53 152	171 868	20,39%	842 795	433,14
Dichomitus squalens	124	7,23%	1 715	63 203	34 171	97 374	11,55%	842 795	468,52
Fomitiporia mediterranea	113	6,59%	1 715	67 681	39 535	107 216	12,72%	842 795	459,16
Fomitopsis pinicola	51	2,97%	1 715	41 438	38 722	80 160	9,51%	842 795	458,31
Ganoderma sp	52	3,03%	1 715	38 088	36 907	74 995	8,90%	842 795	461,70
Gloeophyllum trabeum	45	2,62%	1 715	44 431	44 959	89 390	10,61%	842 795	451,14
Heterobasidion annosum	59	3,44%	1 715	76 158	41 759	117 917	13,99%	842 795	437,73
Laccaria bicolor	52	3,03%	1 715	48 058	47 329	95 387	11,32%	842 795	449,43
Lipomyces starkeyi	509	29,68%	1 715	291 459	56 435	347 894	41,28%	842 795	410,37
Melampsora laricis-populina	81	4,72%	1 715	120 988	71 912	192 900	22,89%	842 795	397,73
Mycosphaerella fijiensis	495	28,86%	1 715	272 953	52 670	325 623	38,64%	842 795	423,91

Continued on next page

Table A.1.4. Summary statistics about missing data information and average gene length for the `genomic46sp_sparse` data set.

Species	MGenes [a]	MGenes (%) [a]	Tgenes [b]	Mdata [c]	Gaps	TMdata[d]	TMdata (%) [d]	Tchars[e]	AGL[f]
Nectria haematococca	529	30,85%	1 715	275 246	53 398	328 644	38,99%	842 795	433,52
Paxillus involutus	85	4,96%	1 715	62 767	46 839	109 606	13,01%	842 795	449,81
Phanerochaete carnosa	220	12,83%	1 715	114 439	36 539	150 978	17,91%	842 795	462,75
Phanerochaete chrysosporium	222	12,94%	1 715	158 917	47 761	206 678	24,52%	842 795	426,07
Phlebia brevispora	66	3,85%	1 715	46 735	40 778	87 513	10,38%	842 795	458,02
Phlebiopsis gigantea	72	4,20%	1 715	67 204	42 870	110 074	13,06%	842 795	445,97
Pleurotus ostreatus	85	4,96%	1 715	72 710	51 094	123 804	14,69%	842 795	441,10
Postia placenta	1 187	69,21%	1 715	619 307	27 518	646 825	76,75%	842 795	371,16
Puccinia graminis f. sp. Tritici	216	12,59%	1 715	134 673	51 984	186 657	22,15%	842 795	437,72
Puccinia triticina	364	21,22%	1 715	210 340	76 437	286 777	34,03%	842 795	411,56
Punctularia strigosozonata	146	8,51%	1 715	88 861	38 601	127 462	15,12%	842 795	455,92
Saitoella complicata	422	24,61%	1 715	254 471	61 422	315 893	37,48%	842 795	407,50
Schizophyllum commune	46	2,68%	1 715	62 313	50 165	112 478	13,35%	842 795	437,58
Serpula lacrymans	162	9,45%	1 715	109 360	40 395	149 755	17,77%	842 795	446,26
Serpula lacrymans S7,9	62	3,62%	1 715	78 880	38 182	117 062	13,89%	842 795	439,04
Sporobolomyces roseus	366	21,34%	1 715	247 623	43 484	291 107	34,54%	842 795	408,96
Stereum hirsutum	89	5,19%	1 715	57 720	34 356	92 076	10,93%	842 795	461,70
Trametes versicolor	104	6,06%	1 715	53 419	34 935	88 354	10,48%	842 795	468,31
Tremella mesenterica	209	12,19%	1 715	135 691	57 922	193 613	22,97%	842 795	431,06
Ustilago maydis	242	14,11%	1 715	109 932	45 682	155 614	18,46%	842 795	466,52
Wallemia sebi	303	17,67%	1 715	195 553	77 947	273 500	32,45%	842 795	403,18
Wolfiporia cocos	57	3,32%	1 715	45 608	37 967	83 575	9,92%	842 795	457,91
Yarrowia lipolytica	524	30,55%	1 715	283 951	65 785	349 736	41,50%	842 795	413,99

Continued on next page

Table A.1.4. Summary statistics about missing data information and average gene length for the `genomic46sp_sparse` data set.

Species	MGenes [a]	MGenes (%) [a]	Tgenes [b]	Mdata [c]	Gaps	TMdata[d]	TMdata (%) [d]	Tchars[e]	AGL[f]
Average	165,50	9,65%		104 715,00	46 769,50	143 270,00	17,00%		441,83
Standard deviation	220,83	12,88%		114 781,27	11 595,34	118 169,51	14,02%		24,03

[a] Missing genes

[b] Total genes

[c] Missing data (in base pairs)

[d] Total missing data (in base pairs)

[e] Total characters (in base pairs)

[f] Average gene length (in base pairs)

Table A.1.5. Summary statistics about missing data information and average gene length for the `combined67sp_sparse` data set.

Species	MGenes [a]	MGenes (%) [a]	Tgenes [b]	Mdata [c]	Gaps	TMdata[d]	TMdata (%) [d]	Tchars[e]	AGL[f]
Agaricus bisporus	179	10,44%	1715	74 190	17 444	91 634	11,82%	775565	445,27
Armillaria tabescens	1 687	98,37%	1715	771 671	810	772 481	99,60%	775565	110,14
Aspergillus carbonarius	560	32,65%	1715	274 794	30 453	305 247	39,36%	775565	407,20
Auricularia delicate	131	7,64%	1715	63 535	27 546	91 081	11,74%	775565	432,12
Bjerkandera adusta	50	2,92%	1715	28 101	11 807	39 908	5,15%	775565	441,84
Botrytis cinerea	540	31,49%	1715	282 823	48 296	331 119	42,69%	775565	378,25
Ceriporiopsis subvermispora	242	14,11%	1715	86 475	9 033	95 508	12,31%	775565	461,68
Cochliobolus heterostrophus	504	29,39%	1715	240 351	33 608	273 959	35,32%	775565	414,21
Colletotrichum higginsianum	509	29,68%	1715	278 417	37 325	315 742	40,71%	775565	381,28
Coniophora puteana	165	9,62%	1715	76 251	13 478	89 729	11,57%	775565	442,47
Coprinopsis cinerea	67	3,91%	1715	32 748	20 091	52 839	6,81%	775565	438,55
Cryptococcus gattii	249	14,52%	1715	123 481	42 887	166 368	21,45%	775565	415,55
Cryptococcus laurentii	1 163	67,81%	1715	668 113	7 792	675 905	87,15%	775565	180,54
Cryptococcus neoformans	213	12,42%	1715	106 521	33 100	139 621	18,00%	775565	423,40
Dacryopinax sp	166	9,68%	1715	97 357	27 963	125 320	16,16%	775565	419,78
Dichomitus squalens	124	7,23%	1715	52 202	8 819	61 021	7,87%	775565	449,12
Fomitiporia mediterranea	113	6,59%	1715	54 827	13 326	68 153	8,79%	775565	441,58
Fomitopsis palustris	1 499	87,41%	1715	745 318	1 850	747 168	96,34%	775565	131,47
Fomitopsis pinicola	51	2,97%	1715	31 157	10 519	41 676	5,37%	775565	441,04
Ganoderma lucidum	787	45,89%	1715	524 399	22 937	547 336	70,57%	775565	245,94

Continued on next page

Table A.1.5. Summary statistics about missing data information and average gene length for the `combined67sp_sparse` data set.

Species	MGenes [a]	MGenes (%) [a]	Tgenes [b]	Mdata [c]	Gaps	TMdata[d]	TMdata (%) [d]	Tchars[e]	AGL[f]
Ganoderma sp	52	3,03%	1715	29 419	9 886	39 305	5,07%	775565	442,73
Gloeophyllum trabeum	45	2,62%	1715	34 085	14 560	48 645	6,27%	775565	435,28
Hemileia vastatrix	1 077	62,80%	1715	645 097	9 264	654 361	84,37%	775565	189,97
Heterobasidion annosum	59	3,44%	1715	56 443	14 953	71 396	9,21%	775565	425,22
Laccaria bicolor	52	3,03%	1715	35 001	16 529	51 530	6,64%	775565	435,38
Lentinula edodes	928	54,11%	1715	612 365	24 402	636 767	82,10%	775565	176,36
Leucosporidium scottii	1 077	62,80%	1715	619 527	20 223	639 750	82,49%	775565	212,88
Lipomyces starkeyi	509	29,68%	1715	258 686	37 022	295 708	38,13%	775565	397,89
Melampsora laricis-populina	81	4,72%	1715	100 968	47 543	148 511	19,15%	775565	383,75
Melampsora medusae f. sp. deltoidis	1 572	91,66%	1715	753 727	1 135	754 862	97,33%	775565	144,78
Melampsora medusae f. sp. tremuloidis	1 568	91,43%	1715	756 958	908	757 866	97,72%	775565	120,40
Melampsora occidentalis	1 479	86,24%	1715	737 870	1 968	739 838	95,39%	775565	151,39
Microbotryum violaceum	839	48,92%	1715	608 175	24 756	632 931	81,61%	775565	162,82
Moniliophthora perniciosa	1 531	89,27%	1715	754 023	2 156	756 179	97,50%	775565	105,36
Mycosphaerella fijiensis	495	28,86%	1715	242 514	35 063	277 577	35,79%	775565	408,19
Nectria haematococca	529	30,85%	1715	246 074	35 462	281 536	36,30%	775565	416,55
Paxillus involutus	85	4,96%	1715	51 530	17 358	68 888	8,88%	775565	433,54
Phakopsora pachyrhizi	1 138	66,36%	1715	659 002	17 365	676 367	87,21%	775565	171,92
Phanerochaete carnosa	220	12,83%	1715	95 797	12 039	107 836	13,90%	775565	446,64
Phanerochaete chrysosporium	187	10,90%	1715	274 311	27 570	301 881	38,92%	775565	310,00

Continued on next page

Table A.1.5. Summary statistics about missing data information and average gene length for the `combined67sp_sparse` data set.

Species	MGenes [a]	MGenes (%) [a]	Tgenes [b]	Mdata [c]	Gaps	TMdata[d]	TMdata (%) [d]	Tchars[e]	AGL[f]
Phellinidium sulphurascens	1 517	88,45%	1715	742 212	1 248	743 460	95,86%	775565	162,15
Phlebia brevispora	66	3,85%	1715	36 402	12 130	48 532	6,26%	775565	440,89
Phlebiopsis gigantea	72	4,20%	1715	50 102	15 515	65 617	8,46%	775565	432,10
Pisolithus microcarpus	1 624	94,69%	1715	764 564	635	765 199	98,66%	775565	113,91
Pisolithus tinctorius	1 641	95,69%	1715	765 481	726	766 207	98,79%	775565	126,46
Pleurotus ostreatus	85	4,96%	1715	55 065	21 168	76 233	9,83%	775565	429,04
Postia placenta	1 187	69,21%	1715	566 205	19 638	585 843	75,54%	775565	359,32
Puccinia graminis f. sp. Tritici	216	12,59%	1715	117 032	32 760	149 792	19,31%	775565	417,46
Puccinia striiformis f. sp. tritici	1 556	90,73%	1715	755 449	1 031	756 480	97,54%	775565	120,03
Puccinia triticina	364	21,22%	1715	187 308	55 962	243 270	31,37%	775565	394,00
Punctularia strigosozonata	146	8,51%	1715	70 737	13 839	84 576	10,91%	775565	440,40
Saitoella complicata	422	24,61%	1715	223 751	40 433	264 184	34,06%	775565	395,50
Schizophyllum commune	46	2,68%	1715	46 126	20 704	66 830	8,62%	775565	424,65
Serpula lacrymans	162	9,45%	1715	88 493	13 938	102 431	13,21%	775565	433,44
Serpula lacrymans S7.9	62	3,62%	1715	56 973	11 910	68 883	8,88%	775565	427,51
Sporobolomyces roseus	366	21,34%	1715	213 588	28 212	241 800	31,18%	775565	395,67
Stereum hirsutum	89	5,19%	1715	45 865	10 414	56 279	7,26%	775565	442,37
Suillus luteus	1 635	95,34%	1715	768 026	1 023	769 049	99,16%	775565	81,45
Taiwanofungus camphoratus	1 080	62,97%	1715	641 545	12 172	653 717	84,29%	775565	191,89
Trametes versicolor	104	6,06%	1715	42 539	8 721	51 260	6,61%	775565	449,60

Continued on next page

Table A.1.5. Summary statistics about missing data information and average gene length for the `combined67sp_sparse` data set.

Species	MGenes [a]	MGenes (%) [a]	Tgenes [b]	Mdata [c]	Gaps	TMdata[d]	TMdata (%) [d]	Tchars[e]	AGL[f]
Tremella mesenterica	209	12,19%	1715	113 991	35 029	149 020	19,21%	775565	416,03
Uromyces appendiculatus	1 170	68,22%	1715	680 286	7 048	687 334	88,62%	775565	161,89
Uromyces viciae-fabae	1 631	95,10%	1715	764 186	177	764 363	98,56%	775565	133,36
Ustilago maydis	242	14,11%	1715	97 225	26 307	123 532	15,93%	775565	442,66
Wallemia sebi	303	17,67%	1715	169 729	50 603	220 332	28,41%	775565	393,22
Wolfiporia cocos	57	3,32%	1715	35 332	9 813	45 145	5,82%	775565	440,54
Yarrowia lipolytica	524	30,55%	1715	252 891	46 037	298 928	38,54%	775565	400,20
Average	583,55	34,03%		306 498,60	19 200,58	325 699,18	42,00%		336,39
Standard deviation	571,96	33,35%		286 317,86	14 390,33	280 445,28	36,16%		129,40

[a] Missing genes

[b] Total genes

[c] Missing data (in base pairs)

[d] Total missing data (in base pairs)

[e] Total characters (in base pairs)

[f] Average gene length (in base pairs)

Table A.1.6. Summary statistics about missing data information and average gene length for the `genomic46sp_dense` data set.

Species	MGenes [a]	MGenes (%) [a]	Tgenes [b]	Mdata [c]	Gaps	TMdata[d]	TMdata (%) [d]	Tchars[e]
Agaricus bisporus	5	0,82%	613	10 590	16 266	26 856	7,68%	349678
Aspergillus carbonarius	85	13,87%	613	65 286	21 655	86 941	24,86%	349678
Auricularia delicate	5	0,82%	613	10 605	20 772	31 377	8,97%	349678
Bjerkandera adusta	0	0,00%	613	5 882	13 799	19 681	5,63%	349678
Botrytis cinerea	75	12,23%	613	69 660	30 876	100 536	28,75%	349678
Ceriporiopsis subvermispora	5	0,82%	613	6 906	12 812	19 718	5,64%	349678
Cochliobolus heterostrophus	64	10,44%	613	48 053	24 135	72 188	20,64%	349678
Colletotrichum higginsianum	61	9,95%	613	67 996	24 948	92 944	26,58%	349678
Coniophora puteana	5	0,82%	613	11 559	13 754	25 313	7,24%	349678
Coprinopsis cinerea	3	0,49%	613	5 013	16 958	21 971	6,28%	349678
Cryptococcus gattii	4	0,65%	613	10 297	26 925	37 222	10,64%	349678
Cryptococcus neoformans	0	0,00%	613	5 560	22 388	27 948	7,99%	349678
Dacryopinax sp	6	0,98%	613	16 557	19 406	35 963	10,28%	349678
Dichomitus squalens	3	0,49%	613	5 228	12 325	17 553	5,02%	349678
Fomitiporia mediterranea	2	0,33%	613	8 832	14 344	23 176	6,63%	349678
Fomitopsis pinicola	0	0,00%	613	6 792	13 849	20 641	5,90%	349678
Ganoderma sp	0	0,00%	613	6 533	13 094	19 627	5,61%	349678
Gloeophyllum trabeum	0	0,00%	613	6 513	15 365	21 878	6,26%	349678
Heterobasidion annosum	0	0,00%	613	17 613	14 909	32 522	9,30%	349678
Laccaria bicolor	0	0,00%	613	10 749	15 707	26 456	7,57%	349678
Lipomyces starkeyi	65	10,60%	613	57 008	26 985	83 993	24,02%	349678
Melampsora laricis-populina	2	0,33%	613	28 206	23 347	51 553	14,74%	349678
Mycosphaerella fijiensis	61	9,95%	613	49 043	25 709	74 752	21,38%	349678

Continued on next page

Table A.1.6. Summary statistics about missing data information and average gene length for the `genomic46sp_dense` data set.

Species	MGenes [a]	MGenes (%) [a]	Tgenes [b]	Mdata [c]	Gaps	TMdata[d]	TMdata (%) [d]	Tchars[e]
Nectria haematococca	64	10,44%	613	46 385	25 328	71 713	20,51%	349678
Paxillus involutus	0	0,00%	613	10 537	16 925	27 462	7,85%	349678
Phanerochaete carnosa	7	1,14%	613	18 519	14 095	32 614	9,33%	349678
Phanerochaete chrysosporium	6	0,98%	613	31 769	19 698	51 467	14,72%	349678
Phlebia brevispora	2	0,33%	613	11 293	13 658	24 951	7,14%	349678
Phlebiopsis gigantea	1	0,16%	613	10 367	14 477	24 844	7,10%	349678
Pleurotus ostreatus	0	0,00%	613	14 742	17 482	32 224	9,22%	349678
Postia placenta	345	56,28%	613	227 201	14 145	241 346	69,02%	349678
Puccinia graminis f. sp. Tritici	4	0,65%	613	9 345	20 537	29 882	8,55%	349678
Puccinia triticina	9	1,47%	613	25 598	34 203	59 801	17,10%	349678
Punctularia strigosozonata	5	0,82%	613	11 825	13 385	25 210	7,21%	349678
Saitoella complicata	49	7,99%	613	47 043	27 558	74 601	21,33%	349678
Schizophyllum commune	0	0,00%	613	13 711	17 047	30 758	8,80%	349678
Serpula lacrymans	1	0,16%	613	18 013	14 645	32 658	9,34%	349678
Serpula lacrymans S7.9	0	0,00%	613	17 488	12 751	30 239	8,65%	349678
Sporobolomyces roseus	15	2,45%	613	48 157	20 112	68 269	19,52%	349678
Stereum hirsutum	3	0,49%	613	8 350	12 342	20 692	5,92%	349678
Trametes versicolor	0	0,00%	613	4 114	12 880	16 994	4,86%	349678
Tremella mesenterica	2	0,33%	613	14 936	22 288	37 224	10,65%	349678
Ustilago maydis	11	1,79%	613	13 224	16 198	29 422	8,41%	349678
Wallemia sebi	18	2,94%	613	29 058	29 882	58 940	16,86%	349678
Wolfiporia cocos	0	0,00%	613	6 321	13 117	19 438	5,56%	349678
Yarrowia lipolytica	64	10,44%	613	50 246	31 565	81 811	23,40%	349678

Continued on next page

Table A.1.6. Summary statistics about missing data information and average gene length for the `genomic46sp_dense` data set.

Species	MGenes [a]	MGenes (%) [a]	Tgenes [b]	Mdata [c]	Gaps	TMdata[d]	TMdata (%) [d]	Tchars[e]
Average	23	3,75%		26 493,98	19 014,04	45 508,02	13,01%	
Standard deviation	55	8,94%		35 765,33	6 045,16	37 834,55	10,82%	

[a] Missing genes

[b] Total genes

[c] Missing data (in base pairs)

[d] Total missing data (in base pairs)

[e] Total characters (in base pairs)

[f] Average gene length (in base pairs)

Table A.1.7. Summary statistics about missing data and average gene length for the `combined67sp_dense` data set.

Species	MGenes [a]	MGenes (%) [a]	Tgenes [b]	Mdata [c]	Gaps	TMdata[d]	TMdata (%) [d]	Tchars[e]	AGL[f]
Agaricus bisporus	5	0,81%	614	6 643	5 356	11 999	3,68%	325 689	515,09
Armillaria tabescens	596	97,07%	614	323 319	513	323 832	99,43%	325 689	103,17
Aspergillus carbonarius	86	14,01%	614	58 074	13 418	71 492	21,95%	325 689	481,43
Auricularia delicate	5	0,81%	614	6 968	8 512	15 480	4,75%	325 689	509,37
Bjerkandera adusta	0	0,00%	614	2 841	3 428	6 269	1,92%	325 689	520,23
Botrytis cinerea	76	12,38%	614	63 870	21 941	85 811	26,35%	325 689	445,87
Ceriporiopsis subvermispora	5	0,81%	614	5 105	2 500	7 605	2,34%	325 689	522,31
Cochliobolus heterostrophus	65	10,59%	614	43 241	15 113	58 354	17,92%	325 689	486,95
Colletotrichum higginsianum	62	10,10%	614	61 597	16 438	78 035	23,96%	325 689	448,65
Coniophora puteana	5	0,81%	614	6 680	3 980	10 660	3,27%	325 689	517,29
Coprinopsis cinerea	3	0,49%	614	3 945	5 924	9 869	3,03%	325 689	516,89
Cryptococcus gattii	4	0,65%	614	7 809	17 323	25 132	7,72%	325 689	492,72
Cryptococcus laurentii	280	45,60%	614	257 036	5 567	262 603	80,63%	325 689	188,88
Cryptococcus neoformans	0	0,00%	614	4 231	12 360	16 591	5,09%	325 689	503,42
Dacryopinax sp	6	0,98%	614	11 213	9 448	20 661	6,34%	325 689	501,69
Dichomitus squalens	3	0,49%	614	3 308	2 379	5 687	1,75%	325 689	523,73
Fomitiporia mediterranea	2	0,33%	614	5 083	4 257	9 340	2,87%	325 689	516,91
Fomitopsis palustris	493	80,29%	614	308 077	1 348	309 425	95,01%	325 689	134,41
Fomitopsis pinicola	0	0,00%	614	3 771	3 370	7 141	2,19%	325 689	518,81
Ganoderma lucidum	82	13,36%	614	169 650	15 911	185 561	56,97%	325 689	263,40

Continued on next page

Table A.1.7. Summary statistics about missing data and average gene length for the `combined67sp_dense` data set.

Species	MGenes [a]	MGenes (%) [a]	Tgenes [b]	Mdata [c]	Gaps	TMdata[d]	TMdata (%) [d]	Tchars[e]	AGL[f]
Ganoderma sp	0	0,00%	614	3 610	2 648	6 258	1,92%	325 689	520,25
Gloeophyllum trabeum	0	0,00%	614	3 911	4 341	8 252	2,53%	325 689	517,00
Hemileia vastatrix	251	40,88%	614	246 533	5 698	252 231	77,45%	325 689	202,36
Heterobasidion annosum	0	0,00%	614	11 443	4 428	15 871	4,87%	325 689	504,59
Laccaria bicolor	0	0,00%	614	6 361	4 765	11 126	3,42%	325 689	512,32
Lentinula edodes	170	27,69%	614	224 613	17 199	241 812	74,25%	325 689	188,91
Leucosporidium scottii	226	36,81%	614	225 014	13 418	238 432	73,21%	325 689	224,89
Lipomyces starkeyi	66	10,75%	614	51 111	16 571	67 682	20,78%	325 689	470,82
Melampsora laricis-populina	2	0,33%	614	21 316	14 613	35 929	11,03%	325 689	473,46
Melampsora medusae f. sp. deltoidis	535	87,13%	614	312 449	798	313 247	96,18%	325 689	157,49
Melampsora medusae f. sp. tremuloidis	534	86,97%	614	315 729	224	315 953	97,01%	325 689	121,70
Melampsora occidentalis	487	79,32%	614	303 472	1 257	304 729	93,56%	325 689	165,04
Microbotryum violaceum	139	22,64%	614	227 099	16 100	243 199	74,67%	325 689	173,66
Moniliophthora perniciosa	505	82,25%	614	313 364	760	314 124	96,45%	325 689	106,10
Mycosphaerella fijiensis	62	10,10%	614	43 213	16 516	59 729	18,34%	325 689	481,81
Nectria haematococca	65	10,59%	614	41 526	16 030	57 556	17,67%	325 689	488,40
Paxillus involutus	0	0,00%	614	7 125	6 140	13 265	4,07%	325 689	508,83
Phakopsora pachyrhizi	280	45,60%	614	251 612	12 211	263 823	81,00%	325 689	185,23
Phanerochaete carnosa	7	1,14%	614	13 939	4 111	18 050	5,54%	325 689	506,82
Phanerochaete chrysosporium	1	0,16%	614	129 479	12 645	142 124	43,64%	325 689	299,45

Continued on next page

Table A.1.7. Summary statistics about missing data and average gene length for the `combined67sp_dense` data set.

Species	MGenes [a]	MGenes (%) [a]	Tgenes [b]	Mdata [c]	Gaps	TMdata[d]	TMdata (%) [d]	Tchars[e]	AGL[f]
Phellinidium sulphurascens	508	82,74%	614	306 728	891	307 619	94,45%	325 689	170,47
Phlebia brevispora	2	0,33%	614	7 313	3 512	10 825	3,32%	325 689	514,48
Phlebiopsis gigantea	1	0,16%	614	5 615	4 103	9 718	2,98%	325 689	515,45
Pisolithus microcarpus	566	92,18%	614	319 732	567	320 299	98,35%	325 689	112,29
Pisolithus tinctorius	564	91,86%	614	318 449	682	319 131	97,99%	325 689	131,16
Pleurotus ostreatus	0	0,00%	614	9 041	6 335	15 376	4,72%	325 689	505,40
Postia placenta	345	56,19%	614	207 531	12 267	219 798	67,49%	325 689	393,65
Puccinia graminis f. sp. Tritici	4	0,65%	614	7 637	11 959	19 596	6,02%	325 689	501,79
Puccinia striiformis f. sp. tritici	523	85,18%	614	314 307	659	314 966	96,71%	325 689	117,84
Puccinia triticina	9	1,47%	614	21 731	25 030	46 761	14,36%	325 689	461,04
Punctularia strigosozonata	5	0,81%	614	7 599	3 385	10 984	3,37%	325 689	516,76
Saitoella complicata	50	8,14%	614	40 370	17 729	58 099	17,84%	325 689	474,45
Schizophyllum commune	0	0,00%	614	8 433	5 952	14 385	4,42%	325 689	507,01
Serpula lacrymans	1	0,16%	614	11 460	4 757	16 217	4,98%	325 689	504,85
Serpula lacrymans S7.9	0	0,00%	614	9 935	3 416	13 351	4,10%	325 689	508,69
Sporobolomyces roseus	15	2,44%	614	38 430	12 691	51 121	15,70%	325 689	458,38
Stereum hirsutum	3	0,49%	614	5 091	2 950	8 041	2,47%	325 689	519,88
Suillus luteus	566	92,18%	614	321 638	224	321 862	98,82%	325 689	79,73
Taiwanofungus camphoratus	243	39,58%	614	242 750	7 784	250 534	76,92%	325 689	202,57
Trametes versicolor	0	0,00%	614	2 221	2 591	4 812	1,48%	325 689	522,60

Continued on next page

Table A.1.7. Summary statistics about missing data and average gene length for the `combined67sp_dense` data set.

Species	**MGenes** [a]	**MGenes** (%) [a]	**Tgenes** [b]	**Mdata** [c]	**Gaps**	**TMdata**[d]	**TMdata** (%) [d]	**Tchars**[e]	**AGL**[f]
Tremella mesenterica	2	0,33%	614	10 100	12 330	22 430	6,89%	325 689	495,52
Uromyces appendiculatus	312	50,81%	614	269 908	3 918	273 826	84,08%	325 689	171,73
Uromyces viciae-fabae	563	91,69%	614	318 272	147	318 419	97,77%	325 689	142,55
Ustilago maydis	11	1,79%	614	10 773	8 823	19 596	6,02%	325 689	507,62
Wallemia sebi	18	2,93%	614	23 208	18 066	41 274	12,67%	325 689	477,21
Wolfiporia cocos	0	0,00%	614	3 600	2 748	6 348	1,95%	325 689	520,10
Yarrowia lipolytica	66	10,75%	614	44 139	22 102	66 241	20,34%	325 689	473,45
Average	141,57	23,06%		104349,12	7987,72	112336,84	34,49%		388,46
Standard deviation	204,84	33,36%		126103,44	6548,67	124314,30	38,17%		161,32

[a] Missing genes

[b] Total genes

[c] Missing data (in base pairs)

[d] Total missing data (in base pairs)

[e] Total characters (in base pairs)

[f] Average gene length (in base pairs)

Table A.1.8. Summary statistics about the prevalence of rusts species for each data set.

Data set	4 species	%	3 species	%	2 species	%	1species	%
basidioPAML	NA	NA	642	63,82%	364	36,18%	0	0,00%
basidioPAML_Hv	395	72,88%	147	27,12%	0	0,00%	0	0,00%
genomic46sp_sparse	NA	NA	1250	72,89%	269	15,69%	196	11,43%
combined67sp_sparse	638	37,20%	656	38,25%	262	15,28%	159	9,27%
genomic46sp_dense	NA	NA	599	97,56%	15	2,44%	0	0,00%
combined67sp_dense	363	59,12%	236	38,44%	15	2,44%	0	0,00%

Table A.1.9. Detailed table with the results of the branch-site test for each gene contained in the basidioPAML and basidioPAML_Hv data sets. Accessible online at https://doi.org/10.1371/journal.pone.0143959.s004.

Table A.1.10. Table containing information regarding the results of the functional annotation, positive selection and evolutionary rate tests for each gene contained in the basidioPAML and basidioPAML_Hv sets. Accesible online at https://doi.org/10.1371/journal.pone.0143959.s005.

Table A.1.11. Enrichment analyses overview. Accessible online at https://doi.org/10.1371/journal.pone.0143959.s006.

A.2 Population genomic footprints of host adaptation, introgression and recombination in Coffee Leaf Rust

A.2.1 Tables

Table A.2.1. List of the *H. vastatrix* isolates used in this study.

Sample	Replicate?	Geographic origin	Year	Pathotype	Mutant	Host	Variety/Selection
1427	no	Kenya	1976	v5	No	C. arabica	Geisha
3019	no	Tanzania	2006	v5	No	HDT derivative	TaCRI 2005/49
3536	no	Brazil	2010	v5	No	HDT derivative	NA
3624	no	Timor	2013	v*	no	HDT derivative	Sarchimor
130a	no	Kenya	1957	v3,5	Yes	C. arabica	NA
2377	no	Phillipines	1996	v5,6,7,8,9	no	HDT derivative	BO2
178	no	India	1958	v2,3,5	No	C. arabica x C. liberica derivative	S. 288-23
2994	no	Tanzania	2005	v2,4,5	No	HDT derivative	TaCRI 2005/34
71	no	Mozambique	1954	v?	No	C. racemosa	NA
92	yes	S. Tomé	1956	v?	No	C. liberica	NA
3302	yes	India	2009	v*	No	HDT	832/1
3305	no	India	2009	v*	No	HDT	832/2
292	no	India	1961	v1,2,5	No	C. arabica	S.12 kaffa
264	yes	Central African Republic	1963	v1,4,?	No	C. excelsa	NA
256a	yes	Uganda	1969	v6,?	Yes	C. canephora	NA
999	no	India	1968	v2,4,5,6	No	C. arabica x C. canephora	S. 2090 3/16
2191	no	India	1992	v2,5,6,7,9	No	HDT derivative	Cauvery, Catimor
178a	yes	India	1960	v2,3,4,5	Yes	C. arabica	CIFC 110/5-68
37	no	Ethiopia	1954	v1,5	No	C. arabica	NA
815	no	India	1965	v2,5,6	No	C. arabica x C. canephora	S. 905 1/7
167a	yes	India	1959	v1,2,3,5	Yes	C. arabica	NA
535	yes	Timor	1963	v5,6	No	HDT derivative	NA

Continued on next page

Table A.2.1. List of the *H. vastatrix* isolates used in this study

Sample	Replicate?	Geographic origin	Year	Pathotype	Mutant	Host	Variety/Selection
137a	no	Phillipines	1957	v1,4,5	Yes	C. arabica	NA
292a	no	India	1963	v1,2,4,5	Yes	C.arabica	CIFC HW12/2-32
256	yes	Uganda	1960	v?	No	C. canephora	Uganda 1343/269
166	yes	India	1957	v2,3,5	No	C. arabica x C. liberica derivative	S.353
32	no	Uganda	1953/54	v?	No	C. canephora	NA
221	no	Madagascar	1959	v?	No	C. canephora	NA
394	no	Tanzania	1962	v?	No	C. canephora	NA

v^{*} Pathotype beyong the capacity of differentiators

$v^{?}$ Pathotype generally unable to infect *C. arabica* differentiators

Table A.2.2. Summary of the 11 assemblies of RAD data using **PyRAD** with information on the assembly parameters, error statistics and loci information.

Assembly	Parameters				Error Statistics				Loci information			
	clustering	mindepth	maxSharedH	-max-missing	*Total locus* error	*Partial locus* error	*Allele* error	*SNP* error	Total loci	Variable loci	Loci w/ middle gaps	%
Var1	0.97	5	2	0.5	40,40%	12,52%	4,64%	3,92%	79 066	28 898	14 639	18,51%
				0.8	22,67%	3,10%	4,24%	3,55%				
Var2	0.95	5	2	0.5	38,55%	11,10%	4,81%	4,21%	53 986	21 504	14 139	27,42%
				0.8	22,08%	2,77%	4,36%	3,69%				
Var3	0.90	5	2	0.5	39,40%	10,62%	5,19%	4,48%	36 083	15 134	12 330	36,06%
				0.8	21,98%	2,84%	4,70%	3,95%				
Var4	0.97	10	2	0.5	53,63%	15,33%	5,35%	4,57%	69 223	23 797	12 532	18,99%
				0.8	29,78%	5,33%	5,02%	4,21%				
Var5	0.97	15	2	0.5	62,78%	17,94%	5,46%	4,77%	64 484	21 616	11 211	18,98%
				0.8	33,62%	6,76%	4,92%	4,21%				
Var6	0.85	5	2	0.5	41,27%	11,58%	5,92%	6,22%	29 072	11 988	10 127	37,06%
				0.8	22,00%	2,90%	5,22%	4,90%				
Var7	0.97	5	5	0.5	41,02%	12,04%	6,14%	5,24%	86 850	37 331	16 625	20,02%
				0.8	23,22%	3,12%	5,12%	4,32%				
Var8	0.97	5	10	0.5	39,22%	10,53%	9,61%	7,74%	100 168	49 817	20 447	21,22%
				0.8	23,22%	3,07%	8,07%	6,49%				
Var9	0.80	5	2	0.5	42,36%	12,02%	7,33%	10,63%	24 352	9 917	8 776	38,66%
				0.8	21,81%	2,85%	6,68%	9,01%				
Var10	0.97	10	10	0.5	52,46%	13,91%	9,03%	7,49%	92 212	45 812	18 485	20,98%
				0.8	30,55%	5,37%	7,44%	6,09%				
Var11	0.97	15	10	0.5	62,11%	16,62%	9,05%	7,84%	85 479	42 632	16 031	20,42%
				0.8	33,43%	6,62%	8,03%	6,74%				

Table A.2.3. Summary statistics of read number and total base pairs for each *H. vastatrix* isolate

Sample	Read number	Base pairs	Average base pairs
166_1	1769191	162367998	92
2994	3554852	323491532	91
256_1	6971482	639831897	92
535_1	4362657	400551143	92
3302_1	5723759	525511551	92
92_1	3392935	311651865	92
37	9509086	865326826	91
815	4941547	449680777	91
167a_1	2343542	209146153	89
130a	5379797	470874199	88
3019	5427871	493773782	91
3536	3651768	332310888	91
2191	4857770	442057070	91
178a_1	7175755	652993705	91
256a_1	5309413	483156583	91
178	4573837	416219167	91
256a_2	2616112	240355275	92
92_2	3125470	284417770	91
2377	4251498	390568230	92
167a_2	7179945	653227337	91
394	2354599	216207781	92
32	2344081	208407211	89
535_2	7232913	658195083	91
264_1	5907335	542426000	92
256_2	3135192	287934790	92
137a	8255660	751086322	91
166_2	993233	91266623	92
71	4147074	377383734	91
178a_2	2415973	221794101	92
292a	3899614	354864874	91
264_2	5016711	456394692	91
3302_2	3343873	304292443	91
3305	5742356	522554396	91
292	3316459	301797769	91
221	3853892	354083658	92

Continued on next page

Table A.2.3. Summary statistics of read number and total base pairs for each *H. vastatrix* isolate

Sample	Read number	Base pairs	Average base pairs
1427	5963852	542710491	91
999	4013998	365273818	91
3624	1740634	159952481	92
Average	4484689	408534611	91
Standard deviation	1900370	172923772	1

Table A.2.4. Summary of the 11 assemblies of RAD data using **PyRAD** with information on the SNP diversity statistics for the *C1+C2* and *C3* groups.

Assembly	-max-missing	C3 group						C1+C2 groups					
		SNPs	SNPs (MAF) [a]	Singletons	%	SNPs (MAF + HWE) [b]	%	SNPs	SNPs (MAF) [c]	Singletons	%	SNPs (MAF + HWE) [d]	%
Var1	0.5	7 503	1 563	6 755	90,03%	574	36,72%	12 892	6 789	8 424	65,34%	4480	65,99%
Var1	0.8	5 425	951	4 925	90,78%	407	42,80%	6 390	3 113	3 625	56,73%	1613	51,81%
Var2	0.5	5 860	1 362	5 276	90,03%	460	33,77%	10 816	5 483	7 127	65,89%	3650	66,57%
Var2	0.8	4 091	821	3 698	90,39%	311	37,88%	5 545	2 597	3 167	57,11%	1357	52,25%
Var3	0.5	3 331	1 197	3 061	91,89%	167	13,95%	7 373	4 263	5 019	68,07%	3122	73,23%
Var3	0.8	2 205	630	2 048	92,88%	112	17,78%	3 319	1 597	1 935	58,30%	874	54,73%
Var4	0.5	6 617	1 227	6 086	91,98%	443	36,10%	9 508	4 307	7 201	75,74%	3444	79,96%
Var4	0.8	3 105	488	2 856	91,98%	229	46,93%	2 576	1 095	1 681	65,26%	702	64,11%
Var5	0.5	4 617	811	4 245	91,94%	300	36,99%	6 072	2 396	4 896	80,63%	2069	86,35%
Var5	0.8	1 678	242	1 543	91,95%	126	52,07%	876	346	587	67,01%	236	68,21%
Var6	0.5	2 863	1 350	2 421	84,56%	112	8,30%	6 150	3 993	4 196	68,23%	2938	73,58%
Var6	0.8	1 758	596	1 564	88,96%	84	14,09%	2 541	1 332	1 460	57,46%	699	52,48%
Var7	0.5	9 784	2 359	8 289	84,72%	1279	54,22%	18 778	9 989	10 355	55,14%	7425	74,33%
Var7	0.8	6 645	1 283	5 789	87,12%	723	56,35%	8 878	5 621	4 175	47,03%	2475	44,03%
Var8	0.5	13 313	3 905	10 407	78,17%	2588	66,27%	25 884	15 676	11 948	46,16%	12781	81,53%
Var8	0.8	8 784	1 692	7 605	86,58%	993	58,69%	12 788	7 018	5 017	39,23%	5141	73,25%
Var9	0.5	3 236	1 988	2 404	74,29%	90	4,53%	5 827	4 091	4 137	71,00%	3084	75,38%
Var9	0.8	1 695	733	1 435	84,66%	63	8,59%	2 216	1 162	1 392	62,82%	645	55,51%
Var10	0.5	12 952	3 927	9 923	76,61%	2848	72,52%	22 593	14 248	10 485	46,41%	13044	91,55%
Var10	0.8	5 856	1 028	5 118	87,40%	702	68,29%	6 355	3 718	2 385	37,53%	3145	84,59%
Var11	0.5	11 116	3 458	8 303	74,69%	2597	75,10%	14 572	10 523	6 912	47,43%	4526	43,01%
Var11	0.8	3 488	578	3 066	87,90%	408	70,59%	2 921	1 400	1 247	42,69%	1011	72,21%

[a] Minor Allele Frequency of 0.05

[b] Minor Allele Frequency of 0.05 and in Hardy-Weinberg equilibrium

[c] Minor Allele Frequency of 0.13

[d] Minor Allele Frequency of 0.13 and in Hardy-Weinberg equilibrium

Table A.2.5. Summary of linkage disequilibrium statistics and results of the significance tests for the complete *C3* data set (`Var1_MM50_maf`), after removing putatively introgressed isolates (`Var1_MM50_NoInt_maf`) and after removing the isolates from the incipient *C3* sub group (`Var1_MM50_NoInt_NoV5_maf`).

Data set	Mean D'	Stdev D'	Mean r^2	Stdev r^2	Significant pairs	%
Var1_MM50_maf	0,863771681714814	0,272010961799205	0,129469921545777	0,25057114535464	218115	17,87%
Var1_MM50_NoInt_maf	0,87212429477204	0,293966809920851	0,084542511461638	0,205699022759217	31703	7,73%
Var1_MM50_NoInt_NoV5_maf	0,889302865870868	0,282153176334849	0,125773993934876	0,264370896362735	9024	10,98%

A.2.2 Figures

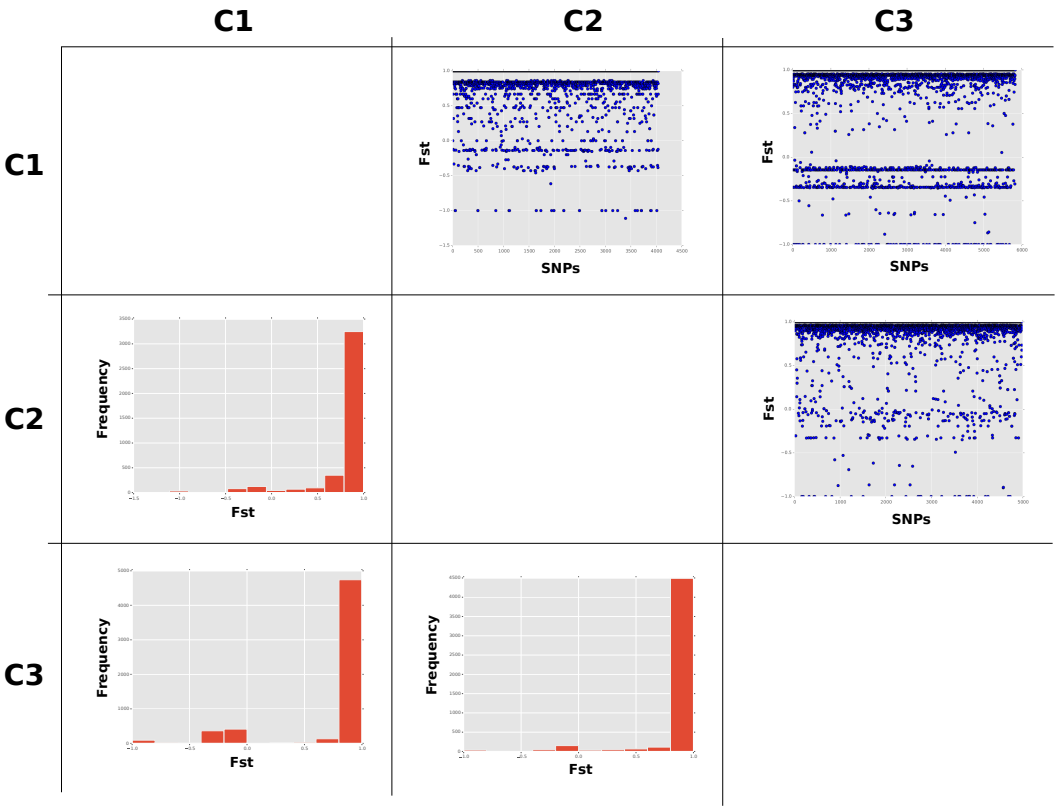

Figure A.1. Triangular matrix with pairwise F_{ST} comparisons between the phylogenetic groups within *H. vastatrix*. Scatter plots in the upper part of the matrix represent the F_{ST} values for each individual SNP segregating between the given group pair. Histograms in the lower part of the matrix represent the distribution of F_{ST} values for the same segregating SNPs.

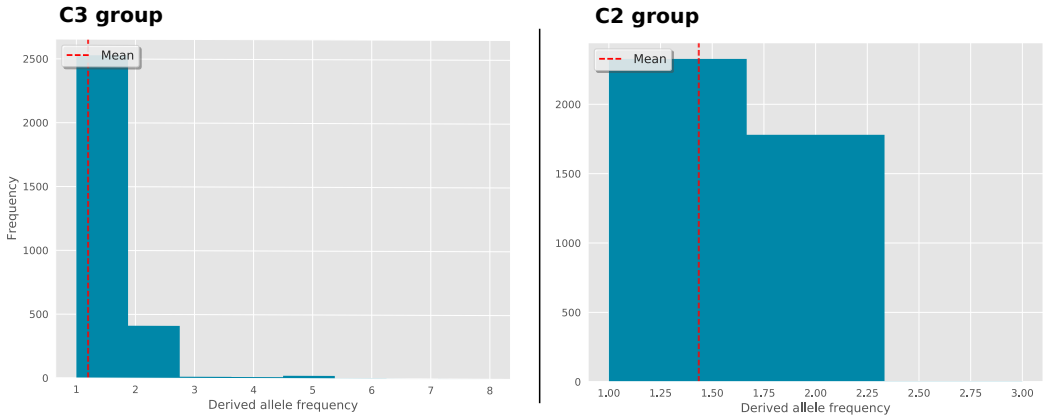

Figure A.2. Allele frequency spectrum for the SNPs of the *C3* phylogenetic group (left) and the *C2* group (right).The vertical dashed line represents the mean of the data set.

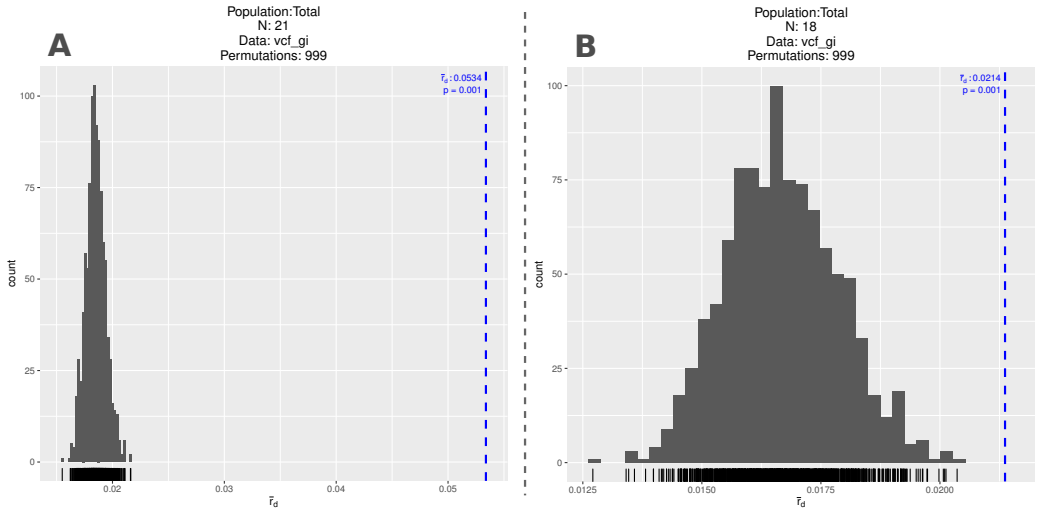

Figure A.3. Results from the Index of Association analysis, using the standardized form ($\bar{r}_d$), for the isolates of the complete *C3* group (A) and after removing putative introgressed isolates (B). The histogram depicts the distribution of $\bar{r}_d$ values expected from unlinked loci. The vertical dashed line represents the observed $\bar{r}_d$ value for the data set.

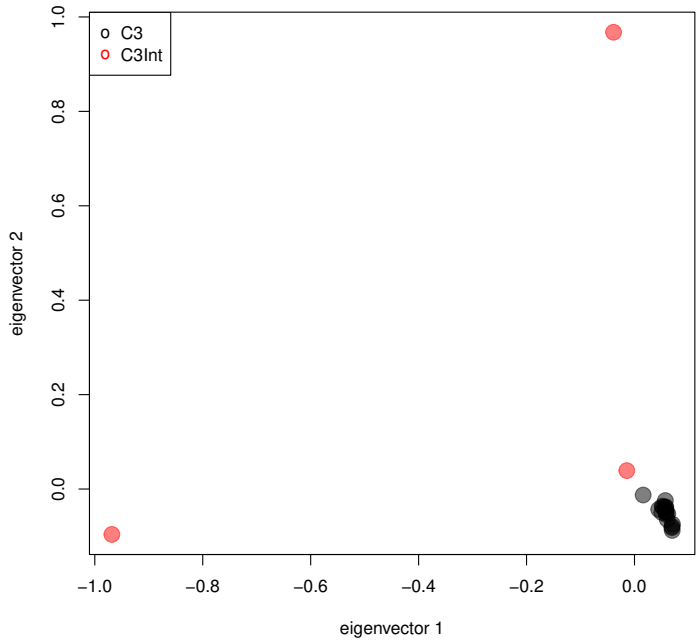

Figure A.4. Principal component analysis of genomic diversity among the 21 *H. vastatrix* isolates from the *C3* group. Isolates are color coded to differentiate the three introgressed isolates from the remaining members of the *C3* group.

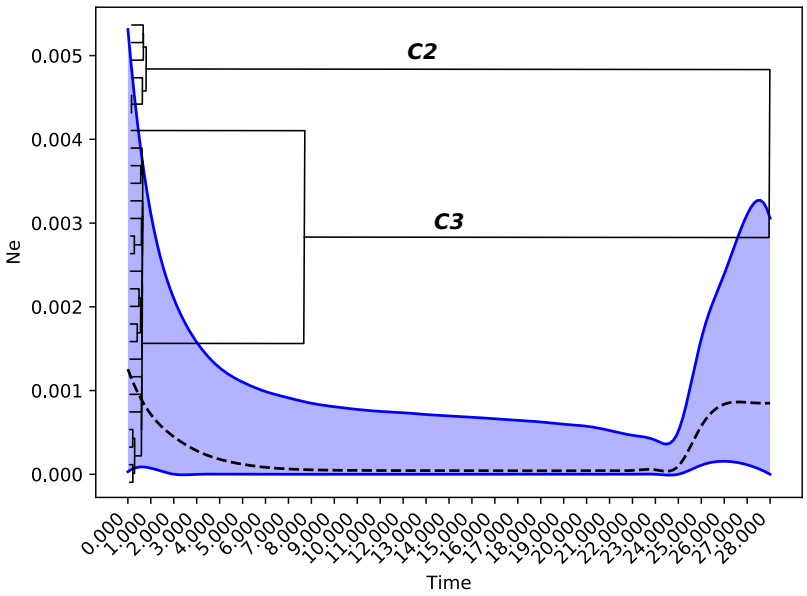

Figure A.5. Extended Bayesian skyline plot depicting the population dynamics of the *C3* group of *H. vastatrix* through time with the inferred phylogeny overlapped. The x-axis is in relative units of time, and the y-axis corresponds to the effective population size. The dashed black line represents the median estimate of the effective population size, while the solid grey lines delimit the 95% high posterior density.

A.3 TriFusion: Streamlining phylogenomic data gathering, processing and visualization

A.3.1 Tables

Table A.3.1. Benchmark of all operations of TriFusion's Process module for test data sets of diverse compositions.

	DS1	DS2	DS3	DS4	DS5	DS6	DS7 [a]	DS8 [b]	DS9	DS10	DS11
Files	614	3093	3093	3093	7378	1	1	1	1	2	52284
Taxa	50	48	141	376	29	40	40	40	376	40	40
Total bases (MB)	17.536	72.385	212.630	567.013	18.405	182.265	182.265	182.265	567.012	364.530	180.801
Parsing											
Alignment reading	0.7 / 5.8	2.7 / 15.3	6.7 / 15.9	18.1 / 17.6	4.0 / 35.5	0.7 / 33.5	18.1 / 27.2	69.2 / 23.6	2.6 / 14.5	1.9 / 33.3	34.4 / 209.2
Main operations											
Conversion (nexus)	0.2 / 0.1	0.7 / 0.2	1.7 / 0.1	4.5 / 0.1	0.8 / 0.3	0.8 / 62.8	0.9 / 62.0	1.0 / 60.9	2.4 / 20.2	1.5 / 52.4	6.4 / 0.5
Conversion (interleave)	2.8 / 0.1	6.8 / 0.3	17.7 / 0.5	45.4 / 2.1	6.1 / 0.4	19.1 / 56.4	18.9 / 56.3	18.8 / 56.3	82.1 / 17.0	38.9 / 52.0	49.9 / 0.8
Concatenation	0.9 / 6.4	3.2 / 25.0	9.0 / 24.8	27.2 / 33.2	3.3 / 13.0	NA / NA	NA / NA	NA / NA	NA / NA	5.0 / 151.3	32.5 / 70.7
Secondary operations											
Collapse	0.5 / 0.1	1.5 / 0.2	2.7 / 0.1	5.4 / 0.1	1.0 / 0.2	1.5 / 61.2	1.5 / 61.2	1.5 / 61.2	3.6 / 21.1	3.0 / 61.5	8.0 / 0.1
Consensus [c]	1.5 / 1.1	6.9 / 0.5	13.6 / 2.3	29.8 / 8.2	1.7 / 0.3	19.1 / 59.7	15.7 / 56.1	15.7 / 53.5	30.9 / 440.1	47.8 / 68.5	54.6 / 2.3
Filter (Taxa contain)	0.2 / 0.3	0.7 / 0.8	1.6 / 0.9	4.1 / 0.9	1.1 / 5.0	NA / NA	NA / NA	NA / NA	NA / NA	0.0 / 0.2	29.1 / 17.9
filter (Minimum taxa; 80%)	0.3 / 0.6	1.1 / 1.0	2.8 / 3.1	7.0 / 8.7	2.6 / 1.5	NA / NA	NA / NA	NA / NA	NA / NA	0.0 / 0.2	48.2 / 20.1
Filter (Missing data; G50; M75)	4.7 / 3.8	17.0 / 13.5	46.3 / 14.1	119.7 / 19.3	9.4 / 8.1	47.6 / 173.2	47.6 / 170.9	48.0 / 170.9	128.3 / 443	101.7 / 115.1	123.7 / 51.9
Filter (Codons; 1 2)	1.9 / 0.2	6.1 / 0.5	20.8 / 0.3	46.5 / 0.4	4.1 / 1.0	NA / NA	NA / NA	NA / NA	NA / NA	31.7 / 142.8	40.8 / 3.4
Filter (Variable; 0, 200)	1.4 / 1.3	6.7 / 1.9	13.3 / 3.8	35.0 / 9.7	3.0 / 4.1	NA / NA	NA / NA	NA / NA	NA / NA	45.6 / 53.6	84.7 / 28.3
Filter (Informative; 0, 200)	2.9 / 1.3	12.6 / 1.9	27.5 / 3.8	64.2 / 10.2	8.1 / 4.0	8.6 / NA	5.8 / NA	9.5 / NA	16.1 / NA	49.5 / 53.9	128.6 / 28.3
Code gaps	3.6 / 6.3	14.4 / 28.3	45.5 / 28.1	116.9 / 27.0	2.3 / 2.9	72.3 / 116.2	64.1 / 116.0	88.0 / 116.2	NA / NA	195.9 / 88.3	19.4 / 11.6
Reverse concatenation											
Reverse concatenation only	0.3 / 6.5	1.4 / 34.8	3.9 / 36.2	9.0 / 34.1	1.6 / 76.0	NA / NA	NA / NA	NA / NA	NA / NA	3.9 / 137.7	14.7 / 424.6
Reverse concatenation + conversion (nexus)	0.4 / 6.5	1.7 / 34.8	4.8 / 36.4	11.6 / 38.2	1.8 / 76.0	NA / NA	NA / NA	NA / NA	NA / NA	5.6 / 139.2	17.1 / 426.7

[a] Interleave (Phylip)

[b] Interleave (Nexus)

[c] Soft mask / single file

Table A.3.2. Comparative benchmark between TriFusion and other existing tools for conversion of phylogenomic datasets

Data set	Files	Taxa	Total bases (MB)		TriFusion	FASconCAT-G	BuddySuite	SequenceMatrix	PGDSpider
DS1	614	48	16,835	Time (s)	1,39	51,57	275,48	NA	NA
				Memory (MB)	60	58	37	NA	NA
DS2	3093	48	72,385	Time (s)	4,49	178,41	1416,15	NA	NA
				Memory (MB)	72	168	38	NA	NA
DS3	3093	141	212,630	Time (s)	9,97	553,44	1366,71	NA	NA
				Memory (MB)	72	327	42	NA	NA
DS4	3093	376	567,013	Time (s)	24,53	1595,44	1435,5	NA	NA
				Memory (MB)	74	710	48	NA	NA
DS5	6186	376	238,638	Time (s)	48,03	2699,97	3101,86	NA	NA
				Memory (MB)	96	1354	37	NA	NA
DS6	52285	40	180,802	Time (s)	62,12	541,39			
				Memory (MB)	296	505			
DS7	1	40	182,225	Time (s)	1,36	†[a]	66,44	9,12	6,99
				Memory (MB)	120	†[a]	2439	1869	2974
DS8	1	376	567,013	Time (s)	3,9	†[a]	227,09	29,82	18,69
				Memory (MB)	81	†[a]	6710	5270	8673
DS9	1	376	1134,026	Time (s)	8,25	†[a]	†[a]	54,44	†[a]
				Memory (MB)	103	†[a]	†[a]	7156	†[a]
DS10	2	40	364,531	Time (s)	2,8	†[a]	NA	NA	NA
				Memory (MB)	133	†[a]	NA	NA	NA

[a] Program crashed due to lack of RAM (Last peak memory usage: 14 000Mb)

Table A.3.3. Comparative benchmark between TriFusion and other existing tools for concatenation of phylogenomic datasets

Data set	Files	Taxa	Total bases (MB)		TriFusion	SCaFoS	FASconCAT-G	SequenceMatrix	BuddySuite
DS1	614	50	17,536	Time (s)	1,56	10,6	118,86	†[c]	6,87
				Memory (MB)	88	111	1522	†[c]	303
DS2	3093	48	72,385	Time (s)	5,58	†[a]	446,58	†[c]	40,25
				Memory (MB)	124	†[a]	6369	†[c]	1171
DS3	3093	141	212,630	Time (s)	12,56	†[a]	†[b]	†[c]	145,82
				Memory (MB)	145	†[a]	†[b]	†[c]	3078
DS4	3093	376	567,013	Time (s)	31,57	†[a]	†[b]	†[c]	376,15
				Memory (MB)	212	†[a]	†[b]	†[c]	7926
DS5	9280	47	212,630	Time (s)	17,01	†[a]	†[b]	†[c]	255,9
				Memory (MB)	223	†[a]	†[b]	†[c]	3430
DS6	2	40	364,531	Time (s)	5,67	201,92	†[b]	15,23*	144,22
				Memory (MB)	248	2970	†[b]	2200	4876
DS7	6186	376	1134,026	Time (s)	70,72	†[a]	†[b]	†[c]	†[c]
				Memory (MB)	351	†[a]	†[b]	†[c]	†[c]
DS8	52285	40	180,802	Time (s)	77,71	2605	†[b]	†[c]	3151,56
				Memory (MB)	469	2976	†[b]	†[c]	4772

[a] Program crashed

[b] Program crashed due to lack of RAM (Last peak memory usage: 14 000Mb)

[c] Program terminated after 60 minutes timeout